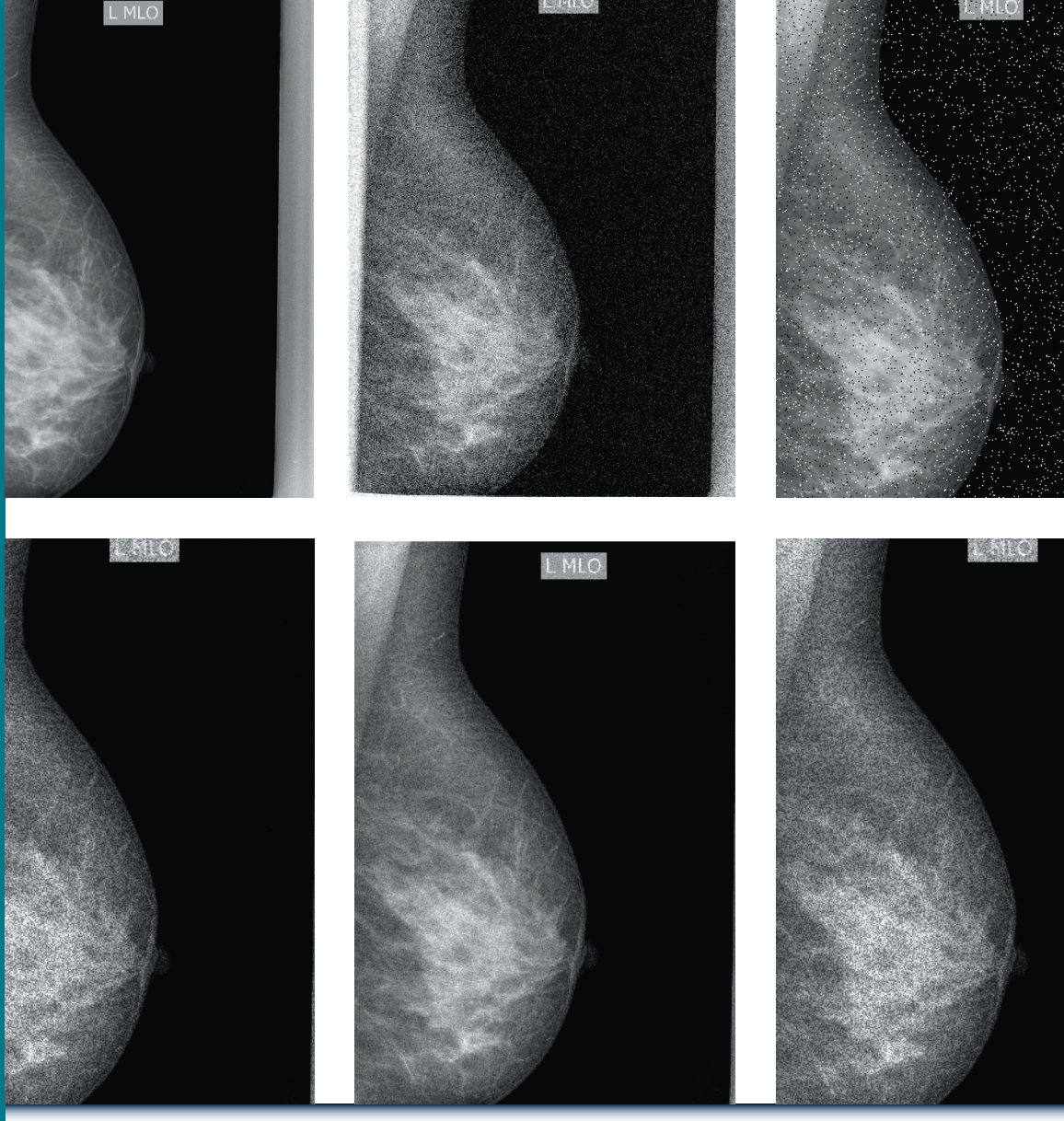


Recuperación de información por contenido orientada a la clasificación de grupos de microcalcificaciones en mamografías - PROTOCAM

Julian David Echeverry Correa
Álvaro Ángel Orozco Gutiérrez
David Augusto Cárdenas Peña
Santiago Marín Mejía

Universidad Tecnológica de Pereira
Editorial UTP

Julián David Echeverry Correa, (Pereira, Risaralda , Colombia,1981)

Doctor en ingeniería por la Universidad Politécnica de Madrid , España , Magister en Ingeniería Eléctrica por la Universidad Tecnológica de Pereira; Ingeniero Electrónico por la Universidad Nacional de Colombia.

Profesor asociado, Facultad de Ingenierías.

Autor del libro "Contributions to speech analytics based on speech recognition and topic identification" (2016). Ha publicado artículos en revistas especializadas nacionales e internacionales.

Miembro del grupo de Investigación en Análisis de Datos y Sociología Computacional

jde@utp.edu.co

Álvaro Ángel Orozco Gutiérrez,
(La virginia, Risaralda, Colombia, 1963)

Doctor en bioingeniería, universidad politécnica de valencia, españa; ingeniero eléctrico, universidad tecnológica de pereira

Profesor titular, facultad de ingenierías

Autor del libro: Vibración Identificación En Línea De Modos Tempranos De Fallas Dinámicas En Máquinas Rotativas, (2011)

Miembro del Grupo de Investigación en AUTOMÁTICA

aaog@utp.edu.co

Recuperación de información por contenido orientada a la clasificación de grupos de microcalcificaciones en mamografías - PROTOCAM

Julian David Echeverry Correa
Álvaro Ángel Orozco Gutiérrez
David Augusto Cárdenas Peña
Santiago Marín Mejía

Universidad Tecnológica de Pereira

Facultad de Ingenierías
Colección Trabajos de Investigación
2020

Recuperación de información por contenido orientada a la clasificación de grupos de microcalcificaciones en mamografías – Protocam / Julian David Echeverry Correa y otros. -- Pereira : Editorial Universidad Tecnológica de Pereira, 2020
81 páginas. -- (Colección Trabajos de investigación).

e-ISBN: 978-958-722-517-4

1. Mamografía – Radiología digital 2. Radiología médica - Métodos de simulación 3. Mamografía – Procesamiento de imágenes 4. Mamografías – Sistemas automáticos.

CDD. 618.190754

Recuperación de información por contenido orientada a la clasificación de grupos de microcalcificaciones en mamografías - PROTOCAM

Publicación financiada con recursos de la Vicerrectoría de Investigaciones , Innovación y Extensión de la Universidad Tecnológica de Pereira

eISBN: 978-958-722-517-4

Proyectos de Investigación:

"Prototipo de un sistema de recuperación de información por contenido orientado a la clasificación de grupos de microcalcificaciones en mamografías". Proyecto financiado por la vicerrectoría de Investigaciones, Innovación y Extensión de la Universidad Tecnológica de Pereira dentro de la convocatoria interna para la financiación de proyectos de grupos, año 2018.

“Prototipo de un sistema de recuperación de información por contenido orientado a la clasificación de grupos de microcalcificaciones en mamografías”, de la Convocatoria de Minciencias Invitación para la conformación de un banco de propuestas elegibles, para el fortalecimiento de programas y proyectos de investigación en ciencias médicas y de la salud, con talento joven e impacto regional año 2018

Universidad Tecnológica de Pereira
Vicerrectoría de Investigaciones, Innovación y Extensión
Editorial Universidad Tecnológica de Pereira
Pereira, Colombia

Coordinador editorial:
Luis Miguel Vargas Valencia
luismvargas@utp.edu.co
Teléfono 313 7381
Edificio 9, Biblioteca Central “Jorge Roa Martínez”
Cra. 27 No. 10-02 Los Álamos, Pereira, Colombia
www.utp.edu.co

Montaje y producción:
María Alejandra Henao Jiménez
Universidad Tecnológica de Pereira
Pereira

Agradecimientos

Este libro fue financiado por la convocatoria de Colciencias -Invitación para la conformación de un banco de propuestas elegibles, para el fortalecimiento de programas y proyectos de investigación en ciencias médicas y de la salud, con talento joven e impacto regional año 2018-, como parte de la Estrategia de Apropiación Social de Conocimiento.

El proyecto que dio origen a este libro lleva por título “Prototipo de un sistema de recuperación de información por contenido orientado a la clasificación de grupos de microcalcificaciones en mamografías”, y fue desarrollado de manera conjunta por el Grupo de Investigación en Automática y por el GIADSc (Grupo de Investigación en Análisis de Datos y Sociología Computacional), ambos grupos adscritos al Programa de Ingeniería Eléctrica de la Facultad de Ingenierías de la Universidad Tecnológica de Pereira. Este proyecto fue financiado por la Vicerrectoría de Investigaciones, Innovación y Extensión de la Universidad Tecnológica de Pereira dentro de la convocatoria interna para la financiación de proyectos de grupos, año 2018.

CONTENIDO

Índice de figuras

Índice de cuadros

Presentación

En este libro se presentan los alcances y resultados del proyecto de investigación denominado "Prototipo de un sistema de recuperación de información por contenido orientado a la clasificación de grupos de microcalcificaciones en mamografías", proyecto al que se ha denominado por el acrónimo **PROTOCAM**.

Este proyecto fue desarrollado de manera conjunta por el Grupo de Investigación en Automática y por el GIADSc (Grupo de Investigación en Análisis de Datos y Sociología Computacional), ambos grupos adscritos al Programa de Ingeniería Eléctrica de la Facultad de Ingenierías de la Universidad Tecnológica de Pereira.

El objetivo del proyecto de investigación y, por tanto, el tema central que se aborda en este libro es el desarrollo de una metodología de recuperación de información cuyo fin último es asistir a los especialistas médicos en el análisis de imágenes de mamografías digitales y en el posterior descubrimiento de patrones que puedan ser indicadores de la existencia de microcalcificaciones en los tejidos mamarios.

Este desarrollo, enmarcado dentro de las ciencias de la computación y, más específicamente, en el área del aprendizaje de máquina, podría coadyuvar a los especialistas en la detección temprana del cáncer de mama, permitiéndoles acceder a través de un sistema inteligente de recuperación de información a datos históricos de diagnósticos confirmados que estén estrechamente relacionados con las características puntuales del caso bajo estudio. En este libro se presentan en detalle cada una de las técnicas utilizadas en los módulos que componen la metodología, además de su funcionamiento y de los resultados obtenidos sobre bases de datos previamente etiquetadas.

Cabe destacar que el proyecto de investigación del cual nació esta metodología fue financiado por la Vicerrectoría de Investigaciones, Innovación y Extensión de la Universidad Tecnológica de Pereira en la convocatoria interna para financiar proyectos de investigación de grupos año 2018. El proyecto de investigación se enmarcó en la categoría de Investigación Aplicada y el Comité de Bioética de la Universidad Tecnológica de Pereira lo clasificó como una investigación sin riesgo, ya que no se realiza ninguna intervención o modificación intencionada de variables biológicas, fisiológicas, psicológicas o sociales de las personas que participan en el estudio. De hecho, la metodología experimental empleada en el desarrollo de este trabajo se basa en el análisis y procesamiento de imágenes de bases de datos y no implica ningún tipo de estudio en el que se involucren directamente pacientes.

Los códigos empleados en el desarrollo de la metodología y las bases de datos públicas que se emplearon en este proyecto se pueden consultar en el repositorio institucional disponible en el siguiente enlace:

https://drive.google.com/drive/folders/1v4qx2gPAyb-v8x66cMcTAZYn6NNEZIcw.

1 CAPÍTULO UNO

1. Preliminares

En este capítulo inicial se presenta formalmente el proyecto de investigación: en la sección 1.1 se hace una introducción al alcance del proyecto y al contenido de este libro; en la sección 1.2 se plantea el problema y la pregunta de investigación; en la sección 1.3 se presenta la motivación y la justificación de este proyecto; en la sección 1.4 se exponen los objetivos del proyecto; en la sección 1.5 se presenta un glosario con la terminología más relevante manejada en este libro relacionada con el cáncer de mama, las microcalcificaciones y su análisis y procesamiento; y por último, en la sección 1.6 se presentan las bases de datos que se emplearon en el proyecto.

1.1. Introducción

El cáncer es una las principales causas de morbilidad y mortalidad en el mundo en la actualidad. Particularmente en mujeres, tanto en países desarrollados, como en países en desarrollo, el cáncer más frecuente es el cáncer de mama. De acuerdo con la Organización Mundial de la Salud (OMS, 2020) la incidencia de esta enfermedad está en aumento debido, entre otras cosas, a una mayor esperanza de vida, al aumento de la urbanización de las ciudades y a la adopción de modos de vida occidentales.

En Colombia, hoy en día, esta enfermedad se perfila como un problema de salud pública debido a que por su causa anualmente fallecen cerca de 2.200 mujeres, y aparecen alrededor de 7.620 nuevos casos (Pardo-Ramos and Cendales-Duarte, 2015). Teniendo presente que esta tendencia sigue en aumento, se hace imprescindible la implementación de estrategias de detección temprana de cáncer de mama, de modo que los tratamientos puedan iniciarse en la mayor brevedad y se contribuya así a la reducción de la tasa de mortalidad debida a esta enfermedad.

Para diagnosticar el cáncer de mama en una etapa temprana, se ha recurrido a la detección de grupos de microcalcificaciones malignas, a partir del Gold Standard para su visualización conocido como imágenes mamográficas. Para estandarizar el análisis, la evaluación y el manejo de hallazgos mamográficos en este tipo de imágenes, el Colegio Americano de Radiología desarrolló una terminología estandarizada denominada *Breast Imaging Reporting and Data System* (BI-RADS). Esta escala estándar incluye además categorías para los hallazgos, los cuales son clasificados de acuerdo con su probabilidad de malignidad. Según el estándar BI-RADS hay siete categorías (de la 0 a la 6) para clasificar una mamografía (S.J. et al., 2020).

La categoría BI-RADS 0 se da a imágenes en las que la evaluación es incompleta y son necesarios nuevos exámenes diagnósticos, bien sean nuevas tomas de imágenes mamográficas bajo otras condiciones o imágenes de ultrasonido. La categoría BI-RADS 1 se da a imágenes con evaluación negativa en las que no se han encontrado masas, calcificaciones sospechosas o áreas con distorsiones. La categoría BI-RADS 2 se da a imágenes consistentes con hallazgos benignos. La categoría BI-RADS 3 se otorga a imágenes con hallazgos probablemente benignos, en donde el riesgo de malignidad se ubica por debajo del 2 %. Existen criterios muy estrictos para clasificar una imagen dentro de esta categoría y es quizás una de las categorías más complejas de clasificar por parte de los especialistas. La categoría BI-RADS 4 se otorga a imágenes en donde hay hallazgos con anormalidades sospechosas. Esta categoría se subdivide a su vez en subcategorías a, b y c. La subcategoría 4.a se da a imágenes en las que hay hallazgos con probabilidades de malignidad que oscilan entre el 2 y el 10 %. La subcategoría 4.b tiene un rango intermedio de probabilidad de malignidad que va del 10 al 50 %. Y finalmente, la subcategoría 4.c tiene un rango de probabilidad de malignidad del 50 al 95 %. La categoría BI-RADS 5 se da a imágenes con hallazgos con una muy alta probabilidad de malignidad (superior al 95 %). Y por último, la categoría BI-RADS 6 se da a imágenes en las que hay hallazgos patológicamente confirmados como positivos para malignidad.

El diagnóstico de la enfermedad a partir de mamografías de la categoría BI-RADS 3 es particularmete difícil, pues como ya se mencionó, los hallazgos mamográficos pueden tener una probabilidad de malignidad menor al 2 %, lo que podría llevar a que estos sean pasados por alto o que sean causantes de falsos positivos, repercutiendo esto en la realización de exámenes diagnósticos o biopsias innecesarias, que no sólo incrementan los costos del diagnóstico, sino que resultan incómodos y estresantes para los pacientes. Lo expuesto anteriormente introduce además incertidumbre en el diagnóstico, y hace que la interpretación de los hallazgos esté ligada al entrenamiento, experiencia y habilidad del médico radiólogo.

Para reducir los problemas inherentes al diagnóstico de esta enfermedad a partir de imágenes de mamografías, se ha optado por disminuir el ruido, eliminar artefactos y mejorar el contraste de las imágenes bajo estudio, lo que resulta en una imagen de mayor calidad. Así mismo, se han planteado formas automáticas de detectar y clasificar grupos de microcalcificaciones sospechosas, empleando sistemas de aprendizaje de máquina que realcen y discriminen entre los patrones asociados a dichos grupos. Sin embargo, sigue existiendo una marcada desconfianza por parte de la comunidad médica en cuanto al uso de las herramientas mencionadas para mejorar la etapa del diagnóstico, y se han desarrollado pocos trabajos que tengan, como objetivo principal, el desarrollo de sistemas de recuperación de información por contenido y que busquen incrementar la confianza de los especialistas en radiología en el uso de dichos sistemas.

El proyecto PROTOCAM se enfocó en el desarrollo de una metodología de recuperación de información que otorgaría al especialista información de apoyo precisa y relevante cuando se enfrente a imágenes con gran incertidumbre en el diagnóstico (p.ej. imágenes categorizadas como BI-RADS 3). La metodología combina dos sistemas de apoyo al diagnóstico: por un lado, se desarrolló un motor de búsqueda

inteligente apoyado por un sistema de recomendación, que entregaría al médico imágenes ya categorizadas en escala BI-RADS, previamente etiquetadas por varios expertos y con diagnóstico confirmado, de pacientes previos cuyos hallazgos presenten similitudes estadísticamente significativas con los del paciente en estudio. Por otro lado, se implementó un módulo de clasificación automática de imágenes mamográficas basándose en la similitud de las imágenes con bases de datos de pacientes previos con diagnósticos confirmados. En concreto se aplicó un sistema automático de reconocimiento de estructuras con base en la detección de contornos. Esta metodología fue desarrollada por el Grupo de Investigación en Automática, y ya ha sido aplicada previamente en proyectos previos con muy buenos resultados en la segmentación de estructuras nerviosas en imágenes de ultrasonido. Ambos sistemas emplearon herramientas de aprendizaje de máquina orientadas a la clasificación y localización de grupos de microcalcificaciones sospechosas.

Gracias a la combinación de estos dos sistemas el especialista podría realizar un diagnóstico más preciso reduciendo las tasas de falsos positivos y los diagnósticos no concluyentes en etapas prequirúrgicas; de esta forma se reducirían las biopsias innecesarias a la vez que se mejoraría la percepción del paciente sobre el diagnóstico y del propio personal sanitario sobre la utilidad de las herramientas de última generación.

Los avances del proyecto PROTOCAM constituyen un paso importante hacia el desarrollo del país en lo relacionado con herramientas de innovación tecnológica que apoyen el trabajo realizado por los especialistas médicos y está en línea con uno de los objetivos del Plan Decenal de Salud Pública (PDSP) 2012-2021 del Ministerio de Salud y Protección Social de Colombia (Minsalud, 2012): “Cero tolerancia con la morbilidad, la mortalidad y la discapacidad evitables”, al ofrecer asistencia tecnológica en el diagnóstico del cáncer de mama que en una instancia temprana puede ser evitable.

1.2. Planteamiento del problema

En el año 2015, el cáncer ocasionó alrededor de 8.8 millones de muertes en el mundo, con el agravante que las detecciones tardías podrían incrementar esta cifra. Una variación de esta enfermedad, el cáncer de mama, ocasionó en el año 2015 cerca de 571.000 de víctimas mortales a lo largo y ancho del mundo (OMS, 2019). En Colombia se diagnostican cada año cerca de 7.200 casos y mueren alrededor de 2.500 mujeres por esta causa (Minsalud, 2013), además la tendencia sigue en aumento (INC, 2012). Para contrarrestar la tendencia en alza de esta enfermedad en Colombia, el Ministerio de Salud y Protección Social ha formulado y desarrollado un programa de detección temprana de cáncer de mama (INC, 2012), que de manera conjunta con entidades educativas y autoridades municipalidades y regionales, permitió capacitar a profesionales de la salud en la orientación a la detección temprana de cáncer de mama. Un estudio diagnóstico fundamental en esta tarea de detección temprana es la mamografía (INC, 2011).

La efectividad de la mamografía para reducir la mortalidad por esta patología ha sido probada ampliamente en distintos estudios. Por ejemplo, antes del empleo

de las mamografías en el diagnóstico, el 2.1 % de todos los carcinomas de mama (proliferación de células epiteliales ductales malignas (Minsalud, 2013)) que eran tratados, correspondían a Carcinoma Ductal In Situ (Ductal Carcinoma In Situ, DCIS, un tipo de cáncer no-palpable); ahora esta cifra ronda el 20 % siendo diagnosticados en una etapa pre-invasiva (INC, 2011; Menezes et al., 2018).

El Colegio Americano de Radiología en su escala estándar BI-RADS), estableció los protocolos necesarios para la evaluación y manejo de hallazgos mamográficos en términos de características morfológicas como la densidad, forma y tamaño de los tejidos (Machado et al., 2018).

En el caso en el que una imagen mamográfica categorizada en BI-RADS 3 presente estabilidad a lo largo del tiempo, el hallazgo es categorizado como benigno o, por el contrario, si se detecta un cambio sustancial en la morfología y densidad, los especialistas recomiendan hacer una biopsia de los tejidos. Repetir los exámenes diagnósticos es menos costoso si se compara con una biopsia; sin embargo, examinar repetidamente a un paciente puede generar en éste ansiedad, debido a la posibilidad de que los hallazgos no sean concluyentes y que permanezca, de forma latente, un posible diagnóstico de cáncer que no esté siendo tratado. Esta situación indeseable puede llegar a dilatarse en algunos casos hasta por más de dos años, tiempo en el cual la conformidad del paciente disminuye y un diagnóstico no concluyente se vuelve intolerable (McDonald et al., 2017). Por esto, los diagnósticos basados en mamografías categorizadas como BI-RADS 3 son los que más problemas pueden llegar a presentar (McDonald et al., 2017).

Las microcalcificaciones (MCC) son uno de los biomarcadores a los que se les ha asociado una relación directa con el DCIS, haciendo de éstas una característica imperante para el diagnóstico de este tipo de cáncer en una etapa temprana (Li et al., 2018) (cabe destacar que, en el contexto de este libro, el término biomarcador se emplea para describir aquellos marcadores morfológicos presentes en las estructuras de los tejidos humanos y que pueden llegar a ser evidencia de la existencia o no de cáncer de mama). Para visualizar estas MCC se emplean principalmente dos técnicas: La primera se basa en imágenes de ultrasonido, las cuales son consideradas como una alternativa confiable para diferenciar entre un quiste y una lesión mamaria sólida, ayudando así a reducir el número de biopsias innecesarias. Sin embargo, para el caso de las MCC sospechosas, las imágenes de ultrasonido presentan una alta sensibilidad y baja especificidad, del 23 % al 36 % (Thibault et al., 2000), lo que significa que no son fácilmente identificables. Esto ocurre debido a que las MCC se pueden confundir con el patrón intrínseco de ruido Speckle ya que ambos se presentan como puntos brillantes en la imagen. Debido a esto la imagen por ultrasonido no se considera el método más adecuado para evaluar la malignidad de las estructuras detectadas en la imagen (Machado et al., 2014).

La segunda técnica se basa en la mamografía, la cual es una imagen generada a partir de rayos-X, considerada en el estado del arte como el *Gold Standard* para la visualización de hallazgos indicadores de cáncer de mama y de etapas tempranas de cáncer no-palpable. Las mamografías permiten por tanto la detección y posterior análisis de las MCC (Machado et al., 2012; Park et al., 2016; Tan et al., 2015).

La detección de grupos de MCC malignas a partir de mamografías está sujeta, en gran parte, a la experticia del radiólogo, lo cual conlleva en algunos casos al sobrediagnóstico del paciente y al aumento de la tasa de falsos positivos (Minsalud, 2013). Distintos trabajos han empleado técnicas de preprocesamiento de imágenes aplicadas a mamografías con el fin de disminuir el ruido, eliminar artefactos y mejorar el contraste de las estructuras de interés en la imagen, con el fin de que la incertidumbre del diagnóstico se viera reducida (Raha et al., 2017; Singh and Kaur, 2018; Soulami et al., 2019).

Sin embargo, los patrones asociados a grupos de MCC son susceptibles a la interpretación del especialista y, por lo tanto, podrían pasarse por alto fácilmente. Así, en algunos trabajos del estado del arte, se han recurrido a técnicas de detección y clasificación automática de los patrones intrínsecos de los grupos de MCC con alta probabilidad de malignidad, a partir de técnicas de aprendizaje de máquina (Malar et al., 2012; Marrocco et al., 2010; Ren, 2012). Estos sistemas automáticos parten de la caracterización local de la imagen, realzando patrones asociados a las MCC agrupadas de carácter benigno y maligno, que luego pueden ser discriminadas mediante sistemas de aprendizaje. Entre las características usualmente empleadas para este tipo de tarea, se destacan los descriptores de textura y morfológicos (Dheeba and Selvi, 2011; Ren, 2012; Tiedeu et al., 2012). Las técnicas de aprendizaje de máquina que mejores resultados han presentado son: las Máquinas de Vectores de Soporte (Support Vector Machines, SVMs) (Dheeba and Selvi, 2011; Ren, 2012), las Redes Neuronales Artificiales (Artificial Neural Networks, ANNs) (Ren, 2012; Tiedeu et al., 2012), entre otros (Malar et al., 2012; Rouhi et al., 2015; Wang et al., 2012; Wei et al., 2012; Xie et al., 2016).

Los sistemas computacionales de apoyo al diagnóstico (*computer-aided diagnosis* - CAD) han servido tradicionalmente para reportar al radiólogo resultados basados en su capacidad para resaltar, discriminar y predecir la malignidad o benignidad de una imagen mamográfica a partir de sus patrones intrínsecos. No obstante, si el sistema no llega a un consenso entre su resultado y el diagnóstico previo, el sistema no conduce a incrementar el grado de certeza de la valoración médica (Tsochatzidis et al., 2017). Además, y luego de una revisión de los sistemas reportados en el estado del arte, los autores consideran que los sistemas CAD no cuentan con herramientas de visualización que permitan comparar los resultados obtenidos por otros especialistas de forma explícita, y que permitan también realizar un análisis del contorno de las calcificaciones.

Por todo ello, y a pesar de que se ha demostrado, en trabajos previos, que la clasificación de grupos de MCC empleando sistemas automáticos ha tenido resultados considerables en relación con el rendimiento obtenido (Malar et al., 2012; Tiedeu et al., 2012; Wang et al., 2012; Wei et al., 2012), el uso de sistemas que empleen herramientas de visualización no ha alcanzado aún un uso extendido debido a que se han desarrollado pocos trabajos que busquen dotar a los sistemas CAD de características que sirvan de soporte para los especialistas en radiología al emplear dichas herramientas (Tsochatzidis et al., 2017).

Un desarrollo combinado de las técnicas actuales de apoyo al diagnóstico que

permita, no sólo mejorar las tasas de clasificación de MCC malignas, sino también reducir el esfuerzo requerido por los especialistas en los casos difíciles, incrementar la certeza en la toma de decisiones y realizar diagnósticos cada vez más precisos y seguros, sería uno de los primeros pasos para la adopción de tecnologías de aprendizaje de máquina que impacte positivamente en la experiencia del paciente y en el apoyo al diagnóstico clínico. Según lo anterior, en el proyecto del que trata este libro se planteó la siguiente pregunta de investigación: ¿podrá un sistema de apoyo al diagnóstico recuperar información histórica relevante de otros pacientes para clasificar y localizar grupos de MCC sospechosas en mamografías?

1.3. Justificación

Según estudios recientes, el cáncer de mama está entre los tipos más comunes de cáncer que pueden desarrollar las mujeres (Li et al., 2018; Scherer et al., 2016; Tsochatzidis et al., 2017). Cada mujer tiene en promedio un 12 % de probabilidad de desarrollar cáncer de mama durante su vida (Scherer et al., 2016). Por esto, diagnosticar y tratar el cáncer de mama en una etapa temprana es fundamental, dado que un tratamiento oportuno puede reducir sustancialmente la tasa de mortalidad por esta enfermedad (Li et al., 2018; Thibault et al., 2000). Para diagnosticar el cáncer de mama en una etapa temprana, es necesario detectar biomarcadores, en este caso las MCC. Las mamografías son las imágenes por excelencia empleadas para detectar MCC, sin embargo, el diagnóstico de la enfermedad empleando este tipo de datos en la mayoría de las ocasiones no es concluyente, probándose incluso innecesario en muchos casos (Scherer et al., 2016; Thibault et al., 2000). Esto es debido a que, en algunos casos, los mismos biomarcadores mencionados pueden ser ambiguos, dado que pueden indicar la presencia de conjuntos benignos o conjuntos con alta probabilidad de malignidad. Esta situación introduce incertidumbre en el diagnóstico, conduciendo a que se presenten muchos falsos positivos, que a su vez pueden llegar a representar intervenciones quirúrgicas costosas e innecesarias, donde además la salud integral del paciente puede verse perjudicada (Li et al., 2018).

Uno de los objetivos del Plan Decenal de Salud Pública 2012-2021 (PDSP) es Çero tolerancia con la morbilidad, la mortalidad y la discapacidad evitables". De acuerdo con lo estipulado con el PDSP existe una proporción de los casos de mortalidad, morbilidad y discapacidad que pueden clasificarse como evitables teniendo en cuenta la existencia de los servicios de salud y la tecnología necesaria, de modo que su presencia significa fallas en el proceso de atención.

La gran mayoría de cánceres no palpables detectados a partir de calcificaciones son DCIS, de los cuales menos del 20 % son de bajo grado, considerados generalmente como sobrediagnóstico, ya que no impacta la reducción de la mortalidad de los pacientes que lo sufren (Mordang et al., 2018). Sin embargo, en una mamografía categorizada como BI-RADS 3 no es posible para un radiólogo distinguir entre calcificaciones asociadas a un DCIS de bajo grado y formas más agresivas del mismo (grado II o III), que deberían ser detectadas lo más pronto posible (Bijker et al.,

2013; Weigel et al., 2015). Por lo tanto, los radiólogos al identificar biomarcadores sospechosos en imágenes mamográficas tienen como única alternativa realizar biopsias. En estos casos, la interpretación de las MCC está ligada a la experticia del radiólogo. Por esto es necesario incrementar la confiabilidad de los diagnósticos a partir de técnicas no invasivas.

De acuerdo con lo anterior, los sistemas CAD ofrecen una asistencia tecnológica sustancial al especialista, al dotar de mayor certeza su toma de decisiones, y a su vez esclarecer el diagnóstico de una enfermedad evitable, así como reducir la probabilidad de un falso diagnóstico positivo de cáncer, y por lo tanto sus implicaciones (Karssemeijer et al., 2004). El objetivo primordial de estos sistemas es el de reducir el esfuerzo requerido para realizar el diagnóstico y el de disminuir el número de falsos positivos que puedan presentarse en éste (Li et al., 2018; Tsochatzidis et al., 2017). En Sari et al. (2017), se demuestra que los sistemas CAD, en su uso como apoyo al diagnóstico de cáncer de mama, mejoran la sensibilidad de los radiólogos del 73.5 % al 87.4 %. De igual forma, se demuestra que su especificidad mejora, pasando del 31.6 % al 41.9 %. Además, se ha demostrado que un sistema de diagnóstico asistido por computador puede ofrecer mayor certeza al especialista a la hora de realizar un diagnóstico de este tipo, donde hoy en día la experticia y habilidad juegan un papel muy importante (Baker et al., 2010; van Luijt et al., 2013).

Dentro de los aportes al diagnóstico temprano de DCIS, en el proyecto PROTOCAM se desarrolló una metodología de apoyo al diagnóstico empleando herramientas de aprendizaje de máquina. Como ya se mencionó previamente, la metodología se compone de dos módulos; el primero de ellos es un sistema de recomendación basado en técnicas de recuperación de información (Baeza-Yates and Ribeiro-Nieto, 2011; Schütze et al., 2008). Este módulo opera de manera similar a como lo hacen los motores de búsqueda web. En estos sistemas el usuario plantea una consulta (query) y el sistema devuelve todos los recursos que se ajusten a los parámetros de la consulta. En el sistema desarrollado, la consulta o query son las imágenes o datos del paciente bajo diagnóstico; y el resultado del sistema de recuperación de información son todos aquellos datos (imágenes, analíticas, informes, reportes, etc.) de pacientes previos con diagnósticos confirmados, ya sean positivos o negativos, y que además cumplan con criterios de similitud con la consulta realizada, los cuales puedan servir al especialista para mejorar la toma de decisiones. El segundo módulo está orientado a la clasificación automática de las imágenes basada en un esquema clásico de reconocimiento de patrones, en el cual se entrenaron modelos de aprendizaje de máquina que permitan reconocer distintos patrones en los tejidos y estructuras, y que hacen posible clasificar y localizar grupos de MCC (benignas y malignas) en mamografías, permitiendo dotar de mayor certeza la toma de decisiones de los especialistas.

De esta manera, con el sistema, el especialista que lo esté empleando podría comparar de forma visual la muestra bajo estudio con muestras de contenido similar recomendadas por el mismo sistema, y que ya hayan sido previamente diagnosticadas por otros especialistas (Tsochatzidis et al., 2017; Wang et al., 2012; Wei et al., 2012). Así mismo, de forma automática, el sistema tiene la capacidad de clasificar los grupos de MCC de las muestras bajo estudio según su probabilidad de malignidad y provee

al especialista de un análisis con el cual comparar su diagnóstico inicial (Malar et al., 2012; Rouhi et al., 2015; Xie et al., 2016).

Con el sistema desarrollado, se dieron los primeros pasos para que, en un futuro, el accionar del especialista se vea favorecido al disminuir el esfuerzo requerido para el diagnóstico de cáncer en aquellos casos donde los biomarcadores en mamografías introducen incertidumbre. Además, podría ayudar a disminuir el número de falsos positivos y los diagnósticos no concluyentes que hacen que la conformidad del paciente con el servicio disminuya McDonald et al. (2017).

Por todo lo anterior, la motivación para esta investigación es contribuir en la detección temprana de cáncer de mama en Colombia, y en la integración de técnicas de aprendizaje de máquina y de sistemas de recuperación de información basados en imágenes, en el proceso diagnóstico. Los posibles trabajos futuros de esta investigación podrían contribuir a mejorar la experiencia del paciente y podrían reducir la incertidumbre de los especialistas a la hora de realizar un diagnóstico basándose en una imagen mamográfica.

1.4. Objetivo general y objetivos específicos

Los siguientes fueron los objetivos del proyecto PROTOCAM. Más adelante en este informe se ahondará en la descripción de cada uno de ellos y se presentarán los métodos que se siguieron para su alcance y los resultados obtenidos.

1.4.1. Objetivo general

- Desarrollar una metodología de recuperación de información por contenido orientada a la clasificación de grupos de microcalcificaciones en mamografías sospechosas empleando técnicas de aprendizaje de máquina.

1.4.2. Objetivos específicos

1. Implementar técnicas paramétricas de preprocesamiento de mamografías con el fin de disminuir el ruido, eliminar artefactos y mejorar el contraste en las imágenes guiado por una metodología de ajuste óptimo de parámetros basado en índices de mejora de la relación señal a ruido de la imagen.

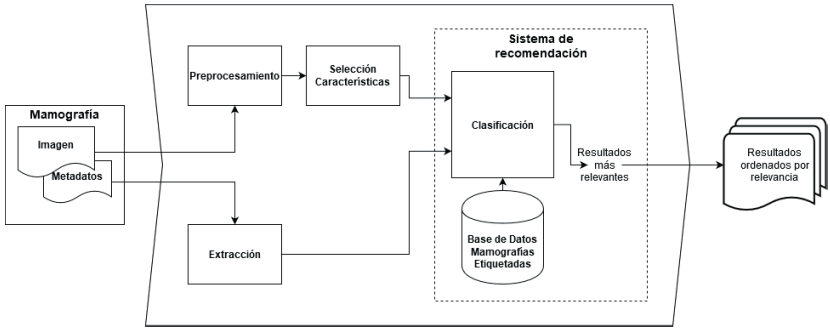
2. Seleccionar características locales relevantes para la clasificación automática, basada en técnicas de aprendizaje de máquina, de grupos de microcalcificaciones en mamografías, que permitan discriminar entre categorías BI-RADS.

3. Desarrollar e implementar una metodología de recuperación de información relevante basada en características de las mamografías y sus metadatos.

En el capítulo 2 se presentarán en detalle las técnicas de preprocesamiento desarrolladas para el alcance del objetivo específico 1. Más adelante, en el capítulo 3, se detalla la metodología empleada en las tareas de clasificación que llevaron al alcance del objetivo específico 2; y finalmente en el capítulo 4 se presentan las técnicas de recuperación de información que condujeron al cumplimiento del objetivo específico 3.

En la Figura 1.1 se presenta un diagrama general que muestra el flujo de la información en el sistema de recuperación de información desde el momento en que se recibe la mamografía como consulta de entrada hasta que se presentan a la salida los resultados más relevantes recuperados de la base de datos. La consulta de entrada se divide en: imagen (matriz de pixeles con valores de intensidad) y metadatos (información relacionada con el caso y/o con la imagen de entrada). La imagen es preprocesada y de ellas se seleccionan las características más relevantes que permitirán posteriormente clasificarla. Tanto los metadatos como las características de la imagen alimentan el módulo de clasificación, el cual recupera de la base de datos aquellas mamografías (ya diagnosticadas y etiquetadas) que sean más relevantes de acuerdo con las características de la imagen de consulta.

Figura 1.1:
Diagrama del flujo de información dentro del marco de trabajo propuesto



1.5. Glosario de términos

A continuación se presentan algunos términos que se emplearán a lo largo de este libro, los cuales están relacionados con el cáncer de mama, las microcalcificaciones y el análisis y procesamiento de los datos.

- **Biomarcador:** Es todo aquel marcador morfológico presente en las estructuras de los tejidos humanos y que puede llegar a ser evidencia de la existencia o no de cáncer de mama.
- **Microcalcificaciones:** se refiere a calcificaciones cuyo diámetro es inferior a 1 mm. Las MCC grandes, de alrededor de un milímetro, son usualmente benignas.

Por el contrario, aquellas cuyo tamaño está bajo los 0.5 mm son usualmente malignas. En términos del número de MCC, se dice que un grupo de 10 o más en un volumen inferior a un centímetro cúbico es más sospechoso que un grupo de 5 de morfología idéntica (Henrot et al., 2014). El cáncer de mama, en etapas tempranas, se manifiesta como MCC, donde su detección puede aumentar las posibilidades de sobrevivencia (Zhang et al., 2014).

- **Mamografía:** La mamografía es una herramienta para el diagnóstico por medio de imágenes radiográficas. Estas son capaces de proveer información clínica notable sin la necesidad de efectuar procedimientos invasivos. Son una técnica esencial en una etapa de diagnóstico inicial, sin embargo, el problema de esta modalidad radica en que al haber varios tejidos en las mamas que exhiben atenuaciones de rayos X similares, las imágenes resultantes presentan un contraste extremadamente bajo entre las diferentes regiones, impidiendo a los radiólogos el reconocimiento efectivo de las características (Russo, 2017).

- **BI-RADS:** Escala desarrollada con el fin de estandarizar la descripción y las recomendaciones para el cuidado de las anomalías detectadas en mamografías, y como resultado, facilitar la comunicación entre radiólogos. De acuerdo con esta escala, la morfología de las MCC puede ser descrita en términos de tres categorías: típicamente benigno (BI-RADS 1-2), riesgo de malignidad intermedio (BI-RADS 3), y alta probabilidad de malignidad (BI-RADS 4-5) (Henrot et al., 2014).

Los siguientes son términos relacionados con el procesamiento digital de imágenes, los sistemas de aprendizaje de máquina y los sistemas de recuperación de información.

- **Preprocesamiento:** es el primer paso en cualquier tipo de procesamiento formal de imágenes. Su función es la de refinar y mejorar la calidad de la imagen de tal forma que ayude a separar los valores de intensidades entre lo que se consideran las Regiones de Interés (*Regions Of Interest* - ROI) y el fondo. Entre los métodos de preprocesamiento se encuentran el filtrado, la eliminación del ruido, el ajuste de contraste, entre otros (Gowri and Amudha, 2014).

- **Caracterización:** consiste en extraer o medir información relevante y no redundante acerca del contenido de una imagen, con el fin de mejorar la representación de esta (Solomon and Breckon, 2011).

- **Aprendizaje de máquina:** es un conjunto de métodos que se usan para detectar automáticamente patrones en datos y que, a partir de estos patrones, pueden predecir datos futuros o realizar otros tipos de toma de decisiones bajo incertidumbre (Robert, 2014). El aprendizaje de máquina se puede dividir en dos tipos: aprendizaje supervisado y aprendizaje no supervisado. En el primer caso, se tiene conocimiento de las variables de entrada y de las salidas del conjunto de entrenamiento. Dependiendo del tipo de valores que tomen las variables de salida (continuas o discretas), se distinguen dos tipos de tareas (Regresión o clasificación). En el aprendizaje no supervisado no se tiene conocimiento sobre

las variables de salida; la tarea principal que se realiza en este tipo de aprendizaje, se conoce como clustering o agrupamiento (Bishop, 2006).

- **Segmentación de imágenes:** es el proceso de dividir una imagen en múltiples segmentos, con el fin de simplificar su representación para facilitar el análisis (Haralick and Shapiro, 1992). Aunque la segmentación de imágenes hace parte de la visión por computador, en ocasiones se realiza a través de algoritmos basados en aprendizaje de máquina, tales como clasificación o agrupamiento (Robert, 2014).

- **Motor de búsqueda de imagen:** es un sistema de retorno de información que ante la entrada de una imagen consulta (*Query*), está diseñado para buscar y devolver imágenes con contenido similar. Usualmente, se requiere de una base de datos de imágenes para que, al analizar y cuantizar una imagen de entrada, el sistema esté en capacidad de retornar las imágenes similares encontradas de la base de datos (Rosebrock, 2017).

- **Sistema de recomendación:** hacen parte de los sistemas de procesamiento de lenguaje natural y son los encargados de filtrar información con el objetivo de entregar al usuario aquellas entradas más relevantes de acuerdo con criterios de búsqueda (Baeza-Yates and Ribeiro-Nieto, 2011; Schütze et al., 2008).

1.6. Materiales

El proyecto de investigación PROTOCAM se clasificó como una investigación sin riesgo, ya que no se realizó ninguna intervención o modificación intencionada de variables biológicas, fisiológicas, psicológicas o sociales de las personas que participan en el estudio. Se contó con bases de datos (BD) públicas y privadas de mamografías. Por una parte, se usaron imágenes de la BD de acceso libre MIAS, la cual contiene 322 imágenes, de los cuales 28 casos corresponden a MCC sospechosas respectivamente, confirmadas por biopsias (Gold Standard). Además, se usaron datos de la BD EJECALS del Grupo de Investigación en Automática, que fue construida en el 2014. Todas las imágenes empleadas cuentan con consentimiento informado, además están anonimizadas para garantizar la privacidad de los pacientes involucrados. Para conformar el grupo de datos a estudiar se respetaron los criterios de inclusión de acuerdo a la Resolución 8430 de 1993 del Ministerio de Salud para la investigación en seres vivos.

BD	Número de imágenes	Tamaño (píxeles)	Formato	Con MCC
MIAS	322	1024×1024	.pgm	28
UTP	510	3560×4640	.tif	49

Adicionalmente cada base de datos cuenta con la siguiente información:

- **ID:** Identificador para cada imagen.

- **Tipo de tejido:** Graso, mixto, glandular.
- **Clase de anormalidad:** Asimetría, masa circunscrita, calcificación, masa especulada, masa mal definida, distorsión de arquitectura y normal.
- **Severidad:** Benigna, Maligna y sin clasificar.
- **BI-RADS:** Clasificación en el estándar BI-RADS (solo en la base de datos UTP).
- **Centro de la anormalidad:** Coordenadas x, y del centro de la anormalidad.
- **Radio:** Radio aproximado del círculo que encierra la anormalidad.

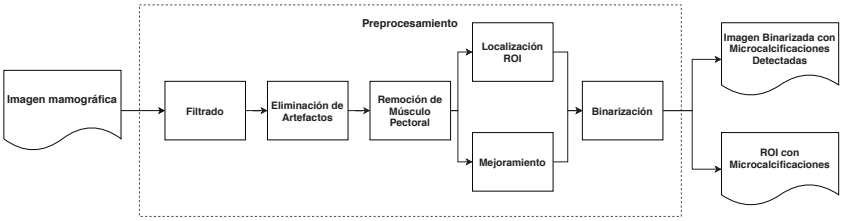
2 CAPÍTULO DOS

2. Preprocesamiento de mamografías para disminuir el ruido, eliminar artefactos y mejorar la imagen

En este capítulo se describen las técnicas empleadas en el procesamiento de mamografías para disminuir el ruido, eliminar artefactos, mejorar el contraste y determinar las características más relevantes de forma local, en términos de la descripción de grupos de microcalcificaciones. Así mismo, se presentará el estado actual de las investigaciones más relacionadas en estas áreas y se presentará una revisión de los resultados más relevantes encontrados en otros proyectos de investigación.

En la sección 2.1 se presenta el marco teórico de las técnicas empleadas en el preprocesamiento de imágenes y la extracción de características. En la sección 2.2 se presenta el estado de la cuestión relacionado con las investigaciones más relevantes que se han realizado en la misma línea de este trabajo. En la sección 2.3 se tratan los métodos usados para llevar a cabo la investigación; y finalmente, en la sección 2.4 se presentan los resultados obtenidos.
En la figura 2.1 se ilustra cómo funciona el módulo de preprocesamiento y sus diferentes etapas.

Figura 2.1:
Diagrama Preprocesamiento



Entrando en detalle respecto al diagrama de flujo de la figura 2.1, la imagen mamográfica ingresa al módulo de preprocesamiento en donde primero se filtra, se

eliminan artefactos presentes en la mamografía y se remueve el músculo pectoral antes de entrar al proceso de localización de microcalcificaciones. Posteriormente se realiza en paralelo la detección de la región de interés y el mejoramiento de la imagen, y por último se binariza la imagen y se entrega una región de interés con microcalcificaciones y una máscara binarizada de esta región de interés.

2.1. Marco teórico

El preprocesamiento de las imágenes de mamografía a menudo tiende a contener tres etapas que son la reducción de ruido o filtrado, eliminación de artefactos y mejoramiento de contraste. Estas etapas son fundamentales de cara a garantizar una correcta y adecuada caracterización de las imágenes (Mehdi et al., 2017; Raha et al., 2017; Singh and Kaur, 2018; Soulami et al., 2019). A continuación en la subsección 2.1.1 se presentan las técnicas de filtrado utilizadas en este proyecto y en la subsección 2.1.2 se trata la eliminación de artefactos.

2.1.1. Filtrado

El ruido en las mamografías se entiende conceptualmente como fluctuaciones aleatorias que afectan el color o el brillo de las imágenes, que se pueden presentar por diferentes factores, en muchos casos como parte del proceso de adquisición de las imágenes. En la etapa de reducción de ruido se implementaron varios tipos de filtros como es el caso del filtro de mediana, mediana híbrido y Gaussiano (George and Sankar, 2017; Mehdi et al., 2017; Saubhagya et al., 2016; Soulami et al., 2019). Los dos primeros ayudan a eliminar efectivamente las líneas verticales y horizontales (conocidas como rasguños) que ocurren en la mayoría de las mamografías y a preservar los bordes (Soulami et al., 2019; Vikhe and Thool, 2016), y el tercero permite atenuar el ruido aditivo.

Filtro de Mediana

El filtro de mediana se utiliza para suavizar las imágenes, reduciendo el cambio brusco de intensidad de un píxel a otro. Es utilizado para reducir el ruido en las imágenes, pero puede resultar en el borrado de los bordes, funciona calculando la mediana de los píxeles vecinos luego de ordenarlos y aplicando esta intensidad a todos los píxeles de la ventana.

Este es el ejemplo de cómo se aplica el filtro de mediana con una ventana de 3x3. En este ejemplo se está considerando el cálculo sobre el píxel central (de intensidad igual a 150):

123	125	126	130	140
122	124	126	127	135
118	120	150	125	134
119	115	119	123	133
111	116	110	120	130

Valores vecinos ordenados: 115,119,120,123,124,125,126,127,150.
Valor de la mediana: 124.

Filtro de Mediana Híbrido

El filtro de Mediana Híbrido es un filtro en ventana de clase no lineal, remueve el ruido impulsivo y preserva mejor los bordes comparado con el filtro de mediana simple, el filtro calcula dos medianas: 1) La mediana de las filas horizontales y verticales 2) La mediana de las diagonales. El valor final es la mediana de esos dos valores y del valor del píxel original (Rakesh et al., 2013).

Filtro Gaussiano

Este filtro es similar al filtro de media pero utiliza una ventana con pesos calculados con una distribución gaussiana de media μ y desviación típica σ:

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Esta sería una ventana 5x5 para un $\sigma = 1{,}0$:

1	4	7	4	1
4	16	26	16	4
7	26	41	26	7
4	16	26	16	4
1	4	7	4	1

Este filtro disminuye el ruido al igual que el filtro de mediana pero preserva mejor los bordes.

2.1.2. Eliminación de Artefactos

Otra etapa del preprocesamiento consiste en los métodos de eliminación de artefactos, los cuales se concentran no sólo en eliminar el rotulado en las imágenes sino también en la eliminación de músculos o tejidos no relacionados (Lbachir et al., 2017; Singh and Kaur, 2018; Soulami et al., 2019). Entre los métodos empleados se encuentran los basados en intensidades, los cuales miden las intensidades entre el músculo pectoral y el tejido mamario y a partir de la diferencia de intensidad eliminan regiones no

deseadas, ya que los niveles de gris del músculo pectoral son más altos que los tejidos aledaños (Lbachir et al., 2017; Singh and Kaur, 2018; Soulami et al., 2019). Otro método es el algoritmo de crecimiento de regiones, el cual tiene un punto semilla de partida y va creciendo al añadir puntos homogéneos similares hasta que los criterios de similaridad y homegeneidad no se cumplan. Un tercer método es el de umbralización cuyo principio es utilizar un valor constante como valor umbral. En este método se compara el nivel de gris de cada píxel de la imagen con dicho umbral, y se reemplaza su valor por un uno si se supera, o por un cero en caso contrario (Lbachir et al., 2017; Vikhe and Thool, 2016).

2.2. Estado de la cuestión

Como etapa previa al proceso de identificación de estructuras y clasificación, las mamografías deben ser preprocesadas. Las etapas de preprocesamiento incluyen: I. Eliminación de ruido, II. Eliminación de artefactos y III. Remoción del músculo pectoral (Nagi et al., 2010a). Estas 3 etapas son de suma importancia debido a que por lo general el músculo pectoral representa la región más predominante y de tejido más denso en las mamografías con proyección MLO (Medio lateral oblicua), y esto puede afectar los resultados del procesamiento de la imagen (Homer and Sickles, 1997), tanto en el nivel de acierto como en la velocidad de análisis dado que se reduciría de forma significante el área a analizar en la imagen (Slavković-Ilić et al., 2016).

Una gran cantidad de algoritmos para la segmentación del seno y músculo pectoral han sido desarrollados, la popularidad de los métodos basados en aprendizaje automático como las redes neuronales convolucionales y el aprendizaje profundo han motivado a varios autores a usar estos métodos para la segmentación de imágenes. Sin embargo, construir un modelo robusto para la segmentación del seno y el músculo pectoral se convierte en una tarea de alta complejidad cuando ambas regiones son bastante similares. Además estos modelos requieren de una gran cantidad de imágenes (e imágenes etiquetadas) para ser construidos y los tiempos de entrenamiento pueden llegar a ser bastantes prolongados, siendo estos algunos de los factores que hacen que la segmentación de mamografías sea una tarea de mediana complejidad (Rampun et al., 2017).

El método de umbralización para la eliminación del músculo pectoral suele dejar cierto solapamiento entre la región de interés y el fondo, resultando inevitablemente en la clasificación errónea de algunos píxeles del fondo como parte del seno y viceversa (Nagi et al., 2010a). Raba et al. (2005) presentaron un método basado en contraste para la segmentación del seno y la extracción del músculo pectoral, pero este enfoque no ha demostrado tener un buen rendimiento cuando se usan imágenes con problemas de fugas de luz en los bordes e inconsistencia en la intensidad de los píxeles en diferentes imágenes (Mustra and Grgic, 2013a). Sun et al. (2006b) usaron una combinación de umbralización adaptativa y estimación de los bordes de los tejidos, calculando la curvatura de la línea del borde del seno, pero este método fracasa en su etapa de realce de contraste para la segmentación del seno debido a los diferentes

contrates de las imágenes mamográficas (Sun et al., 2006a). Otros trabajos como (Chen and Zwiggelaar, 2010) basado en técnicas de crecimiento de regiones falla al trabajar con mamografías donde el músculo pectoral y el seno son homogéneos.

En cuanto a la detección de microcalcificaciones se han realizado diferentes investigaciones en las cuales se ha promovido el uso de técnicas de aprendizaje automático. El-Naqa et al. (2002) utilizaron máquinas de vectores de soporte (SVM) para detectar grupos de microcalcificaciones y lograron una sensibilidad del 94 % a un FP por imagen. Wei et al. (2005) propone un enfoque de aprendizaje Bayesiano conocido como máquinas de vectores de relevancia (RVM). El método es computacionalmente eficiente y proporciona un rendimiento similar en comparación con SVM con una sensibilidad del 90 % a un FP por imagen. Peng et al. (2009) propuso un algoritmo basado en el ruido de resonancia estocástica para detectar microcalcificaciones. Guo et al. (2016) propuso un algoritmo de detección de clúster MC que utiliza la transformación de contorno y la red neuronal acoplada al pulso (PCNN). Chen et al. (2014) propuso un método para la clasificación de grupos de microcalcificaciones utilizando la topología de microcalcificaciones individuales y reportó una precisión de hasta 96 %. Mordang et al. (2016) propuso una técnica semiautomática basada en el aprendizaje profundo para la detección de grupos de microcalcificaciones en los que las regiones de interés o parches de tamaño 13×13 píxeles son manualmente recortadas y se clasifican como grupos de microcalcificaciones o regiones normales. El método logró una sensibilidad cercana al 100 % con una tasa alta de falsos positivos por imagen.

2.3. Métodos

En esta sección se presentan la metodología que se siguió en la ejecución del preprocesamiento de las mamografías previo a la etapa de clasificación y análisis. En la subsección 2.3.1 se describen las técnicas usadas para la parametrización de los filtros utilizados. En la subsección 2.3.2 se habla de las técnicas empleadas para la eliminación de los artefactos presentes en las mamografías. En la subsección 2.3.3 se enseñan los pasos seguidos para la remoción del músculo pectoral que genera ruido en las mamografías y en la subsección 2.3.4 se explica la metodología propuesta para la detección de microcalcificaciones y la extracción automática de las regiones de interés

2.3.1. Filtrado

En el proceso de filtrado se probaron diferentes filtros que fueron puestos a prueba en los tipos de ruido más comunes en las mamografías digitales, a saber, el ruido tipo Gaussiano, el ruido tipo sal y pimienta, el ruido tipo Speckle y el ruido tipo Poisson (Athira et al., 2016). Para seleccionar los parámetros de cada uno de ellos se utilizó una búsqueda exhaustiva en la que se tomaron valores en un rango y se evaluó su desempeño midiendo su Proporción Máxima de Señal a Ruido (**PSNR**), la cual mide la relación de la energía máxima de una señal y el ruido que afecta esta señal. Esta

métrica se utiliza generalmente para medir la calidad de reconstrucción de una imagen de forma cuantitativa.

Ruido adicionado

Para probar los filtros y su desempeño con diferentes parámetros se optó por adicionar los cuatro tipos de ruidos que más afectan las mamografías. De acuerdo con Athira et al. (2016), estos son el ruido tipo Sal y Pimienta, el ruido tipo Speckle, el ruido tipo Poisson y el ruido tipo Gaussiano.

1. **Ruido tipo Sal y Pimienta:** También llamado ruido impulsivo. La principal razón de la aparición de este ruido en las mamografías es el cambio brusco de la señal de la imagen y la presencia de polvo al momento de la adquisición o sobrecalentamiento de componentes defectuosos. Cuando la imagen se ve afectada por este ruido algunos píxeles son reemplazados por valores máximos o mínimos, siendo ’255’ o ’0’ respectivamente.

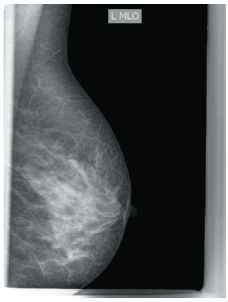
2. **Ruido tipo Speckle:** Este tipo de ruido ocurre cuando la señal de retorno de un objeto causa una interferencia en la adquisición de la imagen causando que la media de grises de un área local aumente.

3. **Ruido tipo Poisson:** Este es un ruido electrónico que es causado por una fluctuación de los rayos X. Es generado por cambios en el número de electrones presentes en los circuitos electrónicos de las máquinas.

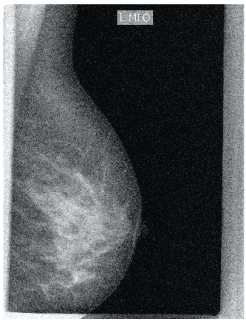
4. **Ruido tipo Gaussiano:** El ruido tipo Gaussiano se distribuye uniformemente sobre toda la señal y sigue la distribución de una campana. Es generado principalmente por ruidos en el circuito eléctrico o el sensor de la máquina.

Para evaluar los filtros y sus capacidades, se agregó cada uno de los cuatro ruidos a una misma mamografía y se aplicó cada uno de los filtros por separado para medir la relación señal a ruido luego de filtradas. La mamografía con la incidencia de los distintos tipos de ruido puede verse en la Figura 2.2.

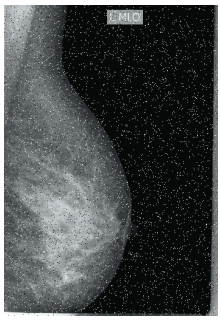
Figura 2.2:
Mamografías con ruido adicionado



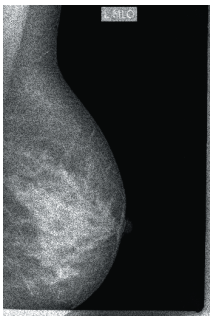
(a) Mamografía original



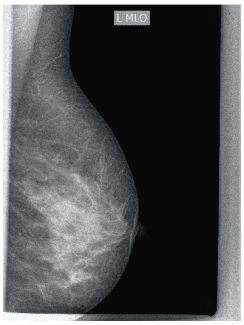
(b) Mamografía con ruido tipo gaussiano con σ= 0,01



(c) Mamografía con ruido tipo sal y pimienta con densidad= 0,05



(d) Mamografía con ruido tipo speckle con varianza= 0,05 y media= 0

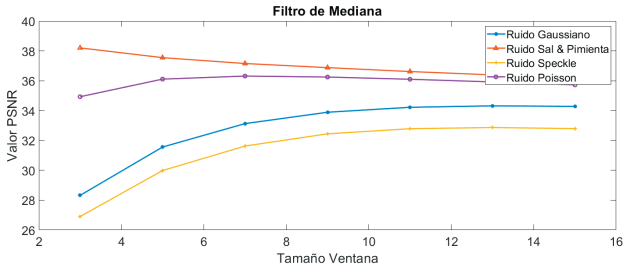


(e) Mamografía con ruido tipo poisson

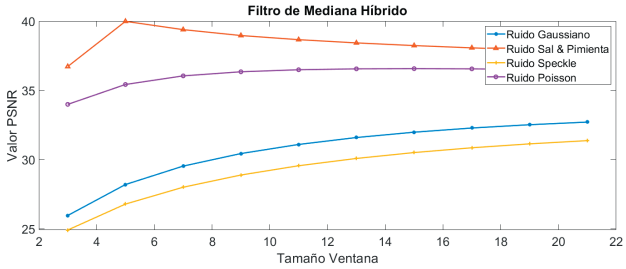
Selección de parámetros

La selección de los parámetros óptimos del filtrado se realizó por medio de una búsqueda exhaustiva en el rango de posibles valores que puede tomar el parámetro de cada uno de los filtros. A la imagen resultante se le midió la relación señal ruido (PSNR) que representa cuantitativamente en escala logarítmica la diferencia entre la imagen original y la imagen reconstruida después del filtrado. En la Figura 2.3 se presentan los resultados experimentales de la sintonización de parámetros.

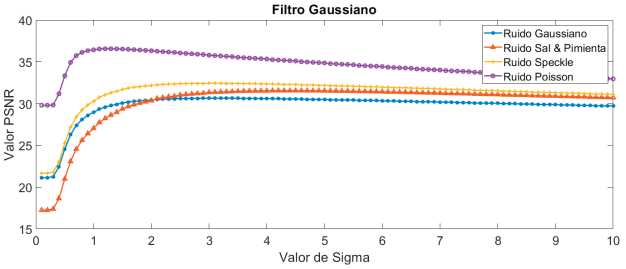
Figura 2.3:
Gráficas de la selección de parámetros



(a) PSNR del filtro de mediana en los diferentes tipos de ruido



(b) PSNR del filtro de mediana híbrido en los diferentes tipos de ruido

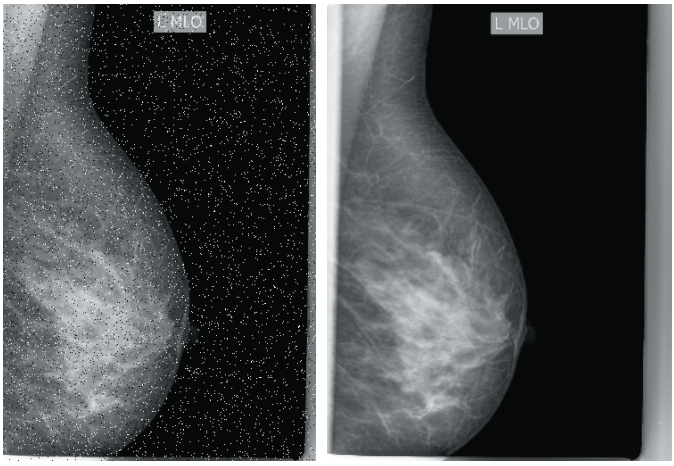


(c) PSNR del filtro gaussiano en los diferentes tipos de ruido

Analizando las gráficas anteriores y probando con los filtros y los parámetros con mejores resultados se tomó la decisión de filtrar las imágenes con un filtro de Mediana con una ventana de tamaño 7×7 para eliminar el ruido producido por la lineas horizontales y verticales. No se utilizó una ventana más grande pues como se evidenció en Gungor and Karagoz (2017), una ventana muy grande puede presentar valores PSNR buenos comparables con ventanas más pequeñas, pero las imágenes se empiezan a volver muy borrosas y se pierde información valiosa; se usó un filtro Gaussiano con un valor de sigma $\sigma = 1$, el filtro de Mediana Híbrido no se utilizó pues los resultados experimentales de la relación señal a ruido evidenciaron un menor desempeño que el

filtro de Mediana simple.
En la Figura 2.4 se muestra con un ejemplo cómo una mamografía afectada con ruido tipo Sal y Pimienta es reconstruida luego de aplicarle los dos filtros parametrizados experimentalmente.

Figura 2.4:
Ejemplo de filtrado



(a) Mamografía con ruido Sal Y Pimienta adicionado

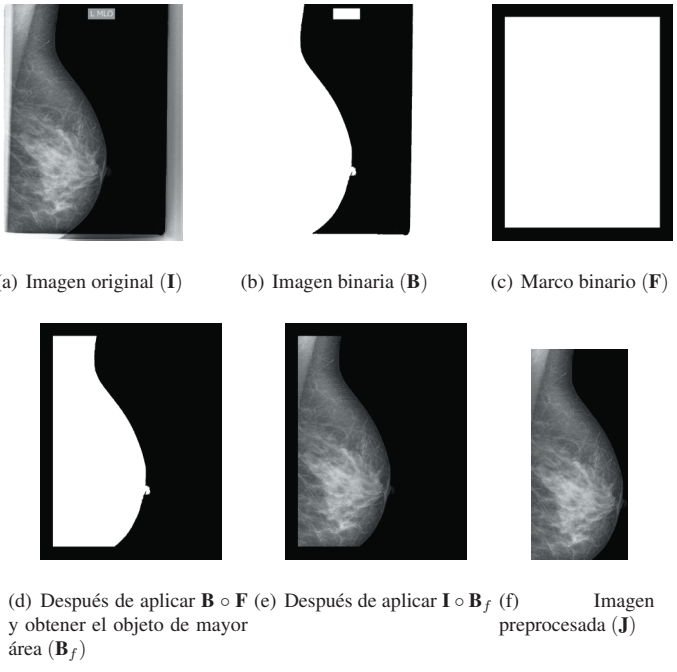
(b) Mamografía filtrada con filtro de Mediana y filtro Gaussiano parametrizados

2.3.2. Eliminación de Artefactos

Una vez eliminado el ruido en la etapa anterior, la mamografía debe pasar por un procesamiento previo en el cual se eliminan todos los artefactos inducidos por el escáner en el momento de digitalización de la imagen (Nagi et al., 2010b). Para esto definimos la imagen filtrada como $\mathbf{I} \in \mathbb{N}^{L\times M}$ y posteriormente usamos el método de umbralización de Otsu para segmentar el la región correspondiente al tejido mamario, el valor de umbral seleccionado fue de $0{,}1$ el cual se obtuvo de forma experimental (Dabass, 2020; Shrivastava and Bharti, 2020). La imagen binaria resultante $\left(\mathbf{B} \in \mathbb{N}^{L\times M}\right)$ se usó para eliminar el marco que en ocasiones rodea el área de interés. Esto se hizo aplicando la operación de multiplicación por elementos $(\circ)$ entre $\mathbf{B}$ y un marco binario $\left(\mathbf{F} \in \mathbb{N}^{L\times M}\right)$ el cual representa un marco estándar cuyo ancho de margen es de $\lceil \frac{M}{14} \rceil$ tal como se se muestra en la Figura 2.5 (c). Finalmente se empleó el método de componentes conectados el cual consiste en identificar las regiones blancas en la figura 2.5 (d), la región más grande (área mayor) representa el área de interés $\left(\mathbf{B}_f \in \mathbb{N}^{L\times M}\right)$ y el resto de las regiones son descartadas; luego se aplica $\mathbf{I} \circ \mathbf{B}_f$ para extraer la región correspondiente al tejido mamario de la imagen original y después se reduce la imagen a un rectángulo que circunscribe el seno (BoundingBox de $\mathbf{B}_f$)

dando como resultado la imagen preprocesada de nuevas dimensiones $\mathbf{J} \in \mathbb{N}^{L' \times M'}$.

Figura 2.5:
Pasos en la eliminación de artefactos.



(a) Imagen original ($\mathbf{I}$) (b) Imagen binaria ($\mathbf{B}$) (c) Marco binario ($\mathbf{F}$)

(d) Después de aplicar $\mathbf{B} \circ \mathbf{F}$ y obtener el objeto de mayor área ($\mathbf{B}_f$) (e) Después de aplicar $\mathbf{I} \circ \mathbf{B}_f$ (f) Imagen preprocesada ($\mathbf{J}$)

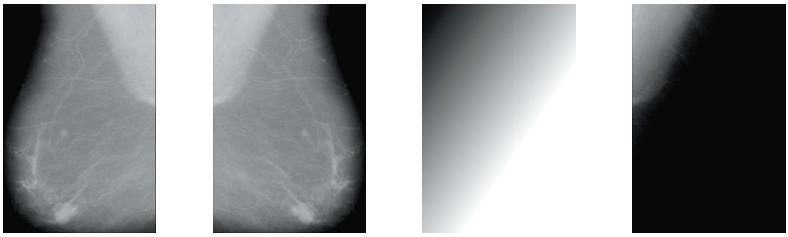
2.3.3. Remoción del Músculo Pectoral

Para eliminar la región correspondiente al músculo pectoral, se implementa una metodología de cuatro pasos: el primer paso es homogeneizar la orientación de las imágenes, es decir, conseguir que todas las imágenes estén orientadas a una misma dirección (bien sea derecha o izquierda). Para eso se divide la imagen preprocesada $\mathbf{J}$ haciendo un corte vertical resultando en dos imágenes $\mathbf{J}_L$ y $\mathbf{J}_R$, cada una de dimensiones $L' \times M'/2$. En cada una de estas imágenes, se cuentan los píxeles distintos de cero. Si el número de píxeles distintos de cero es mayor en la porción $\mathbf{J}_L$, entonces la imagen está orientada a la izquierda; de otro modo la imagen está orientada a la derecha (Shrivastava and Bharti, 2020). En este trabajo se homegeneizaron todas las imágenes con orientación hacia izquierda ($\mathbf{J}_F$) como podemos ver en la Figura 2.6 (b).

En el segundo paso de esta metodología se busca eliminar el fondo de la imagen. Por lo tanto, se deben eliminar todas aquellas estructuras diferentes del músculo pectoral.

Esto se consigue creando una estimación del fondo (**BG**) para luego suprimirla de la imagen usando la siguiente operación $\mathbf{J}_F - \mathbf{BG}$, dando origen a la imagen **P**. La estimación **BG** es un gradiente que va de derecha a izquierda con una orientación de $45°$, tal como se muestra en la figura 2.6 (c). Lo anterior se realiza ya que se se busca preservar el músculo pectoral, el cual se supone está ubicado en el sector izquierdo de la imagen. Se escoge como valor a difuminar $I = \max(\mathbf{J}_F)$ garantizando así la eliminación de las estructuras irrelevantes de más alta intensidad como se puede observar en la Figura 2.6 (d).

Figura 2.6:
Pasos uno y dos de la remoción del músculo pectoral.



(a) Imagen con artefactos removidos (**J**) (b) Imagen orientada a la izquierda ($\mathbf{J}_F$) (c) Estimación del fondo (**BG**), en este caso $I = 251$ (d) Después de aplicar $\mathbf{J}_F - \mathbf{BG}$ removiendo las estructuras diferentes al músculo pectoral (**P**)

El paso tres consiste en la segmentación del músculo pectoral, para esto se utiliza el algoritmo *Region Growing* (RG). La versión estándar de este algoritmo tiene dos parámetros libres: uno de ellos corresponde al punto de inicialización a partir del cual la región segmentada crece; y el otro determina el umbral o nivel de tolerancia (T) a los cambios de intensidad de la región (Adams and Bischof, 1994). Para este algoritmo se escoge como semilla la coordenada $(10, 10)$ la cual se ubica en la parte superior izquierda de la imagen; de esta forma se asegura de que la región segmentada crezca cubriendo la totalidad del músculo pectoral (Raba et al., 2005). El umbral (T) se busca de manera iterativa. Para encontrar un T óptimo, se hace uso del hecho de que el músculo pectoral es una región homogénea en términos de valores similares de los píxeles (Shrivastava et al., 2017). De esta manera, se calcula el valor de uniformidad de la intensidad de la región (UIV) a partir de la siguiente ecuación (Association et al., 2005):

$$UIV = 100\left(1 - \frac{\sigma_P}{\mu_P}\right)$$

Donde μ_P y σ_P son la media y la desviación estándar de la imagen **P** sin considerar los píxeles con valor igual a cero, lo cual garantiza que se está calculando el

UIV para la región pectoral. La búsqueda del umbral empieza con la inicialización $T = \mu_P$ y se actualiza restando $0{,}01$ en cada iteración. Como es de esperar, para cada iteración, el algoritmo RG segmentará una región diferente, por lo tanto se calcula el nuevo valor UIV_{new} para dicha región. El proceso finaliza cuando se haya obtenido un nivel de uniformidad aceptable $UIV_{new} > 70\,\%$ y las variaciones en la uniformidad respecto de la segmentación anterior se mantengan por debajo de un valor predeterminado $|UIV_{old} - UIV_{new}| < 2\,\%$. Los valores de decremento del umbral $(0{,}01)$, el valor de uniformidad mínimo $(70\,\%)$ y la diferencia de uniformidad máxima $(2\,\%)$ fueron escogidos de forma experimental. Como resultado de lo anterior se obtiene la segmentación del músculo pectoral $(\mathbf{B}_p)$. Cabe anotar que en el paso anterior pueden surgir "agujeros" dentro de la región segmentada $(\mathbf{B}_p)$ que hayan quedado como resultado del algoritmo de crecimiento de regiones. Por esta razón se rellenan todos estos agujeros y se aplica una operación morfológica de dilatación con un elemento estructurante cuadrado de tamaño $m \times n$ sobre $\mathbf{B}_p$, esto con el fin de reducir la distorsión y remover píxeles aislados (Nagi et al., 2010b). El tamaño del elemento estructurante se define tomando como referencia el propuesto por Nagi et al. (2010b). El resultado puede verse en la figura 2.7 (a).

El último paso de la metodología de remoción del músculo pectoral es el ajuste del contorno del propio músculo. Para esto se obtiene el perímetro de $\mathbf{B}_p$ y se hace un proceso de limpieza usando la operación morfológica de erosión. El objetivo de la erosión es eliminar las líneas que se ajustan a los límites de la imagen y preservar solo el contorno que divide al músculo pectoral del resto del tejido mamario. Para esto se usa un elemento estructurante de línea cuya longitud debe ser muy menor en comparación con las líneas horizontales y verticales del contorno de la imagen (en este caso se ha escogido experimentalmente un elemento de longitud igual a 10 píxeles) en las orientaciones $90°$ y $0°$.

A continuación, se hace una aproximación del perímetro del músculo pectoral (ver figura 2.8) utilizando un modelo polinomial de tercer grado sobre el borde de la imagen. Se elige una función polinomial cúbica debido a que la forma del músculo pectoral en la mayoría de los casos sigue una curvatura que puede ser mejor aproximada por una función de este orden (Mustra and Grgic, 2013b).

Finalmente, se crea una máscara a partir del contorno corregido $(\mathbf{BF}_p)$ tal como se muestra en la figura 2.7 (c) y se aplica la operación $\mathbf{J}_f \circ \mathbf{BF}_p$ para suprimir el músculo pectoral obteniendo de esta manera la imagen mamográfica preprocesada $(\mathbf{I}_p)$ que se ve en la figura2.7 (d).

Figura 2.7:
Pasos tres y cuatro para la remoción del músculo pectoral.

(a) Segmentación con region growing ($\mathbf{B}_p$)

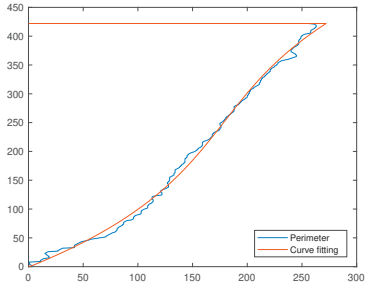
(b) Contorno de $\mathbf{B}_p$

(c) Segmentación del músculo pectoral con la aproximación del perímetro aplicada ($\mathbf{BF}_p$)

(d) Después de aplicar $\mathbf{J}_F \circ \mathbf{BF}_p$ remoción del músculo pectoral ($\mathbf{I}_p$)

En este trabajo se realizaron aproximaciones con funciones de segundo grado y funciones de orden superior, y si bien estas funciones permiten una mayor flexibilidad, también se ven afectadas por la sensibilidad de los parámetros. En la figura 2.8 puede verse la aproximación del contorno del músculo que se consiguió empleando una función cúbica.

Figura 2.8:
Aproximación polinomial del contorno del músculo pectoral.



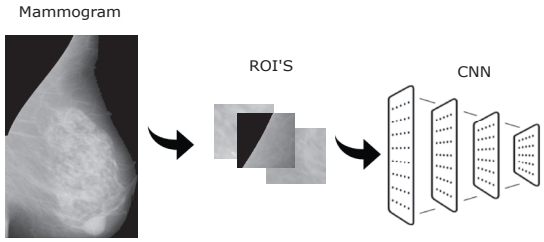
2.3.4. Detección de microcalcificaciones

La metodología propuesta para identificar microcalcificaciones en mamografías consta de dos pasos generales. El primero es la identificación de una potencial región de interés (ROI); en este trabajo la identificación de esta ROI se realizó mediante una red neuronal convolucional (CNN) para clasificar si una ROI específica en la mamografía contiene una microcalcificación o no. El diagrama de este paso se presenta

en la figura 2.9. La ROI específica se ajusta en ventanas de tamaño 101×101 según lo propuesto por Basile et al. (2019), quien manifestó que este tamaño es adecuado para identificar microcalcificaciones sin que la geometría de estas se vea afectada por los efectos de distorsión ocasionados por el pixelado.

En este paso de identificación se utiliza un conjunto de datos (ROI) extraídos manualmente de las bases de datos mini-MIAS y UTP. En estos datos se etiquetan dos clases: "0" para ROI normales (2.360 imágenes) y "1" para ROI con microcalcificaciones (1.932 imágenes), para un total de 4.292 imágenes. Se usa el 80 % (3.500 imágenes) de los datos para el entrenamiento y validación de la red neuronal y el 20 % restante (792 imágenes) para la evaluación. La arquitectura de la CNN consiste en una capa de entrada de tamaño 101×101, ya que se espera que las microcalcificaciones sospechosas estén bien definidas en esta área; 7 capas ocultas con filtros de tamaño 3×3 (Wang and Yang, 2018) y 8, 16, 32, 64, 128, 256 y 512 filtros por cada una de las capas ocultas, cada una de las cuales, tiene asignada una capa de *max-pooling* de tamaño 2×2 la cual permite que las capas extraigan las características de la imagen a mayor escala (Graham, 2014); funciones de activación ReLU (unidad lineal rectificada) en cada capa, la cual es escogida porque estas unidades conducen a un entrenamiento mucho más rápido (Krizhevsky et al., 2012); y la capa de salida es una función softmax, ya que esta permite interpretar la salida como la probabilidad de que una ROI de entrada pertenezca a una determinada clase (Szegedy et al., 2015). El número total de capas fue seleccionado experimentalmente.

Figura 2.9:
Clasificación de la ROI.



Una vez que se ha identificado una ROI que contiene microcalcificaciones ($\mathbf{I}_{ROI}$), se aplica el segundo paso. La idea básica de este paso es mejorar el contraste de las microcalcificaciones para mejorar la precisión de la segmentación. Para esto, cada ROI se procesa en tres etapas: en la primera se elimina el fondo, en la segunda se eliminan las estructuras irrelevantes y en la tercera, se identifican las microcalcificaciones. Para eliminar el fondo, se utiliza una operación morfológica de apertura. Posteriormente se utiliza una reconstrucción wavelet a partir de los coeficientes de aproximación para mejorar el detalle de las microcalcificaciones, y finalmente, la segmentación se realiza mediante un método de binarización. A continuación se presenta en detalle el desarrollo de cada una de estas tres etapas.

Para realizar la primera etapa de procesamiento, se elimina el fondo de la imagen aplicando una transformación morfológica de apertura $\otimes$ a $\mathbf{I}_{ROI}$, usando un elemento estructurante S para determinar la información correspondiente al fondo de la imagen y luego suprimirla de la imagen original usando la siguiente expresión $\mathbf{I}_{ROI} - (\mathbf{I}_{ROI} \otimes S)$, donde S es un disco de radio r, el resultado puede verse en la Figura 2.10 (b).

La segunda etapa es la reconstrucción de $\mathbf{I}_{RBG}$ a partir de los coeficientes wavelet de aproximación W_Φ. Esta es una técnica ampliamente utilizada para mejorar los grupos de microcalcificaciones, ya que características involucradas como la variabilidad, ocurrencia a diferentes escalas y orientaciones, y la caracterización por cambios discontinuos de intensidad, así como variaciones globales más sutiles en la textura. Hacen que el procesamiento wavelet sea extremadamente necesario (Sandino Garzón et al., 2018), el objetivo es aumentar el contraste de las microcalcificaciones incrementando la intensidad en las regiones donde éstas se encuentran, filtrando todo tipo de tejidos y estructuras irrelevantes utilizando la siguiente transformación.

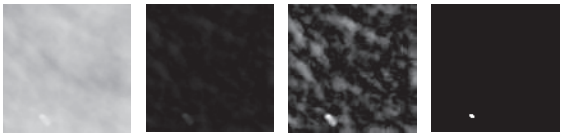
$$W_\Phi = \sum \mathbf{I}_{RBG}\Phi$$

$$\mathbf{I'}_{RBG} = \sum W_\Phi \Phi$$

En esta transformación, se empleó la wavelet (Φ) biortogonal (bior3.5). Esta función base fue seleccionada debido a que ésta ha sido exitosamente usada en muchas aplicaciones cuyo propósito general es detectar los detalles que aparecen a diferentes escalas y mejorarlos selectivamente dentro de diferentes niveles de resolución (Mencattini et al., 2008). Se hizo una descomposición en el nivel 1, y la reconstrucción se realizó en el nivel cero para mantener la resolución de la imagen original, como se puede ver en la Figura 2.10 (c).

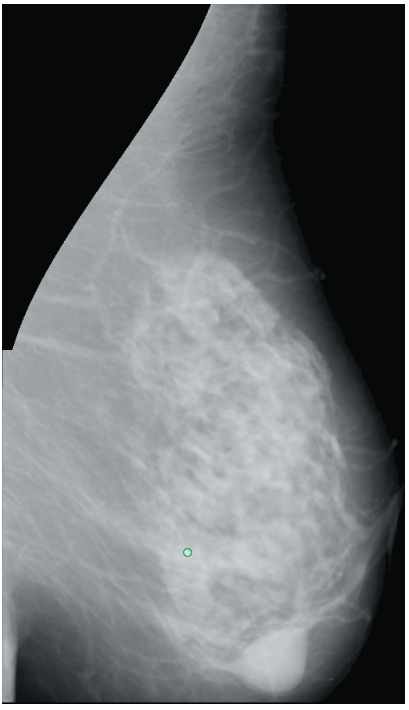
La tercera etapa es la binarización de $\mathbf{I'}_{RBG}$; esta se realizó utilizando la técnica de umbralización de Otsu, obteniendo así el número de objetos (microcalcificaciones) contenidos en $\mathbf{I}_{ROI}$ Figura 2.11 (d), de la imagen binaria resultante se extrae el centroide de todos los objetos, estas serán las coordenadas de las microcalcificaciones en la ROI, como se puede observar en el resultado final mostrado en la Figura 2.11, donde las coordenadas de las MCC dentro del ROI se han convertido a coordenadas globales de la imagen $\mathbf{I}_p$ sumando el desplazamiento horizontal y vertical realizado en el muestreo.

Figura 2.10:
Etapas para la identificación de MCC en la ROI.



(a) ROI original $\mathbf{I}_{ROI}$ (b) Substracción del fondo ($\mathbf{I}_{RBG}$) (c) Después del realce con Wavelets $\mathbf{I'}_{RBG}$ (d) Después de binarizar $\mathbf{I'}_{RBG}$

Figura 2.11:
Resultado final detección de microcalcificaciones.



2.4. Resultados

En esta sección se presentan los resultados obtenidos del preprocesamiento de las mamografías, desde la etapa de eliminación de artefactos hasta la etapa final de binarización. En la evaluación del método propuesto para la eliminación de artefactos y remoción del muscular pectoral, se utiliza la base de datos mini-MIAS de referencia. La base de datos contiene 322 imágenes del mismo tamaño 1024×1024 con 8 bits por píxel. Los resultados de eliminación de artefactos obtenidos mediante inspección visual demuestran una efectividad de $99{,}69\,\%$.

Los resultados de la remoción del músculo pectoral se muestran en la tabla 2.1.

Categoría	Número de Imágenes	Porcentaje
Exitoso	295	91.6150 %
Aceptable	16	4.9690 %
Inaceptable	11	3.4160 %

Cuadro 2.1:
Resultados de la remoción del músculo pectoral sobre la base de datos mini-MIAS

Además, también se usaron 255 imágenes en proyección MLO de la base de datos UTP. Todas estas imágenes tienen un tamaño de 3560×4640. Los resultados de eliminación de artefactos obtenidos mediante inspección visual demuestran una efectividad de $97{,}65\,\%$. sobre el total de 510 imágenes (incluyendo las de proyección CC - cráneo caudal).

Categoría	Número de imágenes	Porcentaje
Exitoso	211	82.7451 %
Aceptable	28	10.9804 %
Inaceptable	16	6.2745 %

Cuadro 2.2:
Resultados de la remoción del músculo pectoral sobre la base de datos UTP

2.5. Conclusión

En este capítulo se presentaron los métodos empleados y los resultados obtenidos en la detección automática de clusters de microcalcificaciones en mamografías con proyección MLO.

Esta parte del sistema consta de cuatro etapas: *i)* un filtrado de la mamografía utilizando sintonización de parámetros por búsqueda exhaustiva basada en la relación señal a ruido; *ii)* eliminación de artefactos usando el método de umbralización de Otsu; *iii)* remoción del músculo pectoral usando una estimación y eliminación del

fondo para eliminar estructuras irrelevantes, en combinación con el algoritmo de segmentación *region growing* usado para identificar el músculo pectoral; *iv)* finalmente la localización de microcalcificaciones, la cual se llevó a cabo usando una red neuronal convolucional para identificar las regiones de la mamografía con potencial de microcalcificaciones, en combinación con una técnica de realce de contraste basada en operaciones morfológicas y reconstrucción wavelet.

Las etapas *i)* y *ii)* conforman el preprocesamiento de la mamografía, los resultados obtenidos mediante inspección visual muestran ser adecuados ya que se evidencia un 99.6 % y 97.6 % de efectividad en la eliminación de artefactos en las bases de datos mini-MIAS y UTP respectivamente, junto con un 91.6 % y 82.7 % de éxito en la eliminación del músculo pectoral en las mismas bases de datos.

Por otro lado, se obtuvo un nivel de acierto del 83.3 % en la evaluación del sistema de identificación de microcalcificaciones para el cual se usaron imágenes de ambas bases de datos.

El buen rendimiento del sistema se debe en gran parte a la novedosa estrategia utilizada para segmentar el músculo pectoral, en donde, por medio de una búsqueda iterativa se encuentra el umbral de segmentación óptimo T. Cabe recordar que esta búsqueda se basa en el hecho de que el músculo pectoral es una región homogénea cuyo grado de homogeneidad le permite al algoritmo RG determinar si el músculo ha sido total o parcialmente segmentado.

Se puede concluir que la aplicación de este algoritmo fue de vital importancia en este proyecto ya que facilita enormemente la búsqueda de ROI sospechosas al remover zonas con componentes de alta intensidad.

3 CAPÍTULO TRES

3. Clasificación automática de grupos de microcalcificaciones en mamografías para discriminar entre categorías BI-RADS.

En este capítulo se describen las técnicas empleadas en la detección y clasificación de microcalcificaciones. Así mismo, se presenta el estado actual de las investigaciones más relacionadas en estas áreas y se presenta una revisión de los resultados más relevantes encontrados en otros proyectos de investigación. En la sección 3.1 se presenta el marco teórico de las técnicas empleadas en la clasificación y localización de microcalcificaciones. En la sección 3.2 se presenta el estado de la cuestión relacionado con las investigaciones más relevantes que se han realizado en la misma línea de este trabajo. En la sección 3.3 se habla de los métodos usados para llevar a cabo la investigación; y finalmente, en la sección 3.4 se presentan los resultados obtenidos.

3.1. Marco teórico

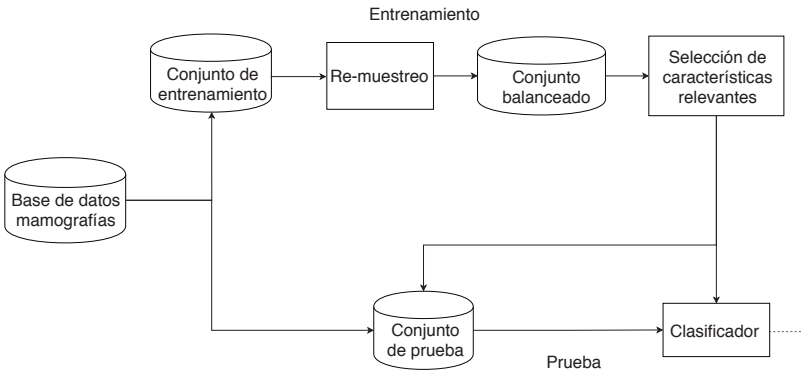
La tarea de clasificación automática de mamografías para la detección de cáncer de mama, tiene como objetivo asignar automáticamente una clase o etiqueta a una mamografía, a partir del entrenamiento de un clasificador con imágenes previamente caracterizadas y etiquetadas certeramente. Esta etiqueta puede ser, benigna o maligna, o la escala BI-RADS correspondiente a cada una. De manera general, un clasificador genera un modelo discriminante el cual asigna una clase a cada nueva mamografía, a partir de la similitud de esta con las usadas anteriormente para su entrenamiento.

En la tarea de clasificación automática de mamografías para la detección de cáncer de mama, existen diversos patrones inusuales, los cuales son difíciles de detectar debido a que estos eventos ocurren con mucha menos frecuencia que los que suceden comúnmente. Este es precisamente el caso de algunas de las microcalcificaciones catalogadas en la BIRAD-4, las cuales ocurren con mucha menor frecuencia que las demás. Sin embargo, no es solo las mamografías en esta escala las que se encuentran sub-representadas. Generalmente, las bases de datos de mamografías suelen presentar una distribución no uniforme entre sus clases (escalas BI-RADS), ya que la frecuencia

de ocurrencia de cada uno de los tipos de anormalidad es completamente diferente. Este desbalance afecta generalmente a la mayoría de algoritmos estándar, por lo cual, técnicas de remuestreo deben ser usadas para contrarrestar estos efectos no deseados.

Por otra parte, la caracterización de las imágenes de mamografías, puede llevar a la obtención de un conjunto con cientos e incluso miles de características, de las cuales muchas podrían no ser necesarias, irrelevantes o redundantes. Esto hace necesario que exista también una etapa de análisis de relevancia donde se seleccione un sub-conjunto de variables que además de disminuir la complejidad del entrenamiento, aumente la capacidad del clasificador para discriminar y reduzca incluso tiempos de cómputo. Por lo tanto, de manera general el diagrama de la etapa de clasificación puede ser resumido en la figura 3.1. Una vez se obtiene la base de datos de mamografías caracterizada, se obtiene una matriz de características con tantas filas como imágenes y tantas columnas como características, y un vector de etiquetas donde se relaciona si la mamografía es benigna o maligna, o la escala BI-RADS a la que pertenece. Seguidamente, esta matriz es separada en dos conjuntos: uno que servirá para entrenar el sistema y otro de menor tamaño par probar el rendimiento de este. Posteriormente, el conjunto de entrenamiento debe ser remuestreado para equilibrar el tamaño de las clases y permitir que el clasificador no se genere sesgos hacia ninguna clase. Después, se deben seleccionar aquellas características que ofrecen mayor explicación de las etiquetas, es decir las más relevantes para la tarea de clasificación y finalmente, con los datos remuestreados y con las características apropiadas se entrena un clasificador que permitirá decir si una nueva imagen es o no maligna o decir a qué escala BI-RADS corresponde.

Figura 3.1:
Diagrama de flujo de la etapa de clasificación



A continuación se enuncian las principales temáticas y métodos relacionados con la clasificación de mamografías asociadas a cáncer. Se describen desde las generalidades de la tarea hasta los métodos particulares que fueron usados en esta investigación.

3.1.1. Clasificación

La clasificación es el problema de establecer una regla para identificar a cuál clase corresponde una nueva observación, basándose en el conjunto de datos de entrenamiento cuyas categorías son conocidas. Los algoritmos de clasificación permiten abstraer la información, llevándola a una representación adecuada para la toma de decisiones.

La clasificación es una rama del aprendizaje de máquina cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. De manera más concreta, se trata de crear metodologías capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos. Una rama del aprendizaje de máquina es el aprendizaje supervisado, el cual busca encontrar una función a partir de un conjunto de datos de entrenamiento, los cuales consisten en pares de objetos: una componente del par son los datos de entrada y el otro, los resultados deseados. El objetivo del aprendizaje supervisado es crear una función que pueda predecir la salida correspondiente a cualquier entrada válida después de haber sido sometido a una serie de datos ejemplos (datos de entrenamiento), es decir, el sistema generaliza con base en los datos que no ha visto. Este tipo de aprendizaje soluciona principalmente tareas de regresión donde la salida son valores numéricos, y de clasificación donde la salida es una etiqueta de clase.

Algunos ejemplos de sistemas de clasificación son etiquetado automático de piezas o producto industrial como correcto o defectuoso (Di Leo et al., 2017), sistemas de seguridad para identificar si una persona tiene acceso o no a cierto lugar (Singla and Sharma, 2019), detección de tumores en rayos-X (Shan et al., 2016), clasificación de pacientes como enfermos o no (Avci and Dogantekin, 2016), clasificación de mamografías asociadas a cáncer, entre otros.

En un escenario de clasificación, una función de discriminación $g : \mathcal{X} \rightarrow \mathcal{Z}$ se aprende a partir de un conjunto $\{\boldsymbol{x}_n, z_n\}_{n=1}^{N}$, donde $\boldsymbol{x}_n \in \mathcal{X} \subseteq \mathbb{R}^P$ es un vector de entrada de características P-dimensional , correspondiente a la n-ésima muestra con etiqueta de salida $z_n \in \mathcal{Z}$. El proceso de clasificación consiste en asignar una clase, z_n, de un conjunto de clases, $\mathcal{Z}$, a cierta instancia, representada por un vector de características o atributos.

3.1.1.1. Máquinas de vectores de soporte

Las máquinas de vectores de soporte (Support Vector Machines, SVMs) son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik en compañía . Las SVMs están propiamente relacionadas con problemas de clasificación y regresión. Específicamente, a partir de un conjunto de muestras de entrenamiento, es posible etiquetar las clases y construir un modelo que prediga la clase de una nueva muestra. De manera general, una SVM es un modelo que representa cada instancia en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación. Cuando las nuevas muestras se proyectan en dicho espacio, en función de los espacios a los que pertenezcan, pueden ser clasificadas

a una u otra clase.

Dicho de otra manera, una SVM genera un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión.

Dicho hiperplano puede ser representado por la siguiente ecuación:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b = \sum_{i=1}^{n} x_i w_i + b = 0$$

Este hiperplano parte el espacio en dos dimensiones. Específicamente, se define una función de mapeo $y = sign(f(\mathbf{x})) \in \{1, -1\}$,

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b = \begin{cases} > 0, & y = sign(f(\mathbf{x})) = 1,\ \mathbf{x} \in P \\ < 0, & y = sign(f(\mathbf{x})) = -1,\ \mathbf{x} \in N \end{cases}$$

Cualquier punto $\mathbf{x} \in P$ en el lado positivo del plano es asignada la etiqueta 1, mientras que cualquier punto $\mathbf{x} \in N$en el lado negativo del plano es asignada la etiqueta -1. Un punto $\mathbf{x}$ de clase desconocida será clasificado como P si $f(\mathbf{x}) > 0$, o como N si $f(\mathbf{x}) < 0$.

En particular, para construir un hiper-plano $\mathbf{x}^T\mathbf{w} + b = 0$ que separe las dos clases $P = \{(\mathbf{x}_i, 1)\}$ y $N = \{(\mathbf{x}_i, -1)\}$, se tiene que satisfacer

$$y_i(\mathbf{x}_i^T\mathbf{w} + b) \geq 0$$

tanto para $\mathbf{x}_i \in P$ como para $\mathbf{x}_i \in N$. Entre todos los planos posibles que satisfacen esta condición, se desea encontrar el óptimo H_0 que sepera las dos clases con el margen máximo (la distancia entre el plano de decisión y los puntos más cercanos. Dicho plano óptimo debería estar en el medio de las dos clases.

El problema de encontrar el hiperplano de separación óptimo en términos de $\mathbf{w}$ y b puede ser formulado como

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}||\mathbf{w}||^2 \qquad \text{(Función objetivo)} \\ \text{sujeto a} & y_i(\mathbf{x}_i^T\mathbf{w} + b) \geq 1, \ \text{ or } \ 1 - y_i(\mathbf{x}_i^T\mathbf{w} + b) \leq 0, \quad (i = 1, \cdots, m) \end{array}$$

Sin embargo, el algoritmo anterior funciona bien solo para datos linealmente separables. Si los datos no son linealmente separables, dichos datos se deben mapear en un espacio de características de más alta dimensión::

$$\mathbf{x} \longrightarrow \phi(\mathbf{x})$$

en el cual las clases pueden ser linealmente separables. La función de decisión en este nuevo espacio se convierte finalmente en:

$$f(\mathbf{x}) = \phi(\mathbf{x})^T\mathbf{w} + b = \sum_{j=1}^{m} \alpha_j y_j(\phi(\mathbf{x})^T\phi(\mathbf{x}_j)) + b$$

3.1.2. Clasificación de datos desbalanceados

En la tarea de reconocimiento de patrones y aprendizaje de máquina existen diversos patrones inusuales, los cuales son difíciles de detectar debido a que estos eventos ocurren con mucha menos frecuencia que los que suceden comúnmente (Maalouf and Trafalis, 2011). Fenómenos tales como desastres naturales, expresiones de genes de cáncer, transacciones fraudulentas, la ocurrencia de piezas defectuosas en un proceso industrial, son fenómenos que ocurren con menos regularidad que sus respectivos casos contrarios. Esta situación induce bases de datos que naturalmente tendrán más observaciones de los casos "normales"que de estos patrones inusuales.

En términos de clasificación, cualquier conjunto de datos que exhiba una distribución desigual entre sus clases puede considerarse desbalanceado. Como se mencionó anteriormente, este fenómeno se produce cuando hay más muestras en una clase que la otra clase en un conjunto de datos de entrenamiento. En un conjunto de datos desbalanceado, la clase mayoritaria tiene un gran porcentaje de todas las muestras, mientras que las instancias de la clase minoritaria sólo ocupan una pequeña parte de las observaciones.

Por lo general, sin consideración del problema del desbalance de clases, un algoritmo de clasificación tenderá a predecir que las muestras desconocidas pertenecen a la clase mayoritaria e ignoran completamente la clase minoritaria. Sin embargo, en muchas de las aplicaciones, la clase minoritaria es de vital importancia. Un ejemplo clásico es la clasificación de pacientes con cáncer o sanos a partir de imágenes de mamografías. Basados en la experiencia, el número de pacientes sanos supera notablemente el número de pacientes con cáncer. En consecuencia, en muchos algoritmos estándar de aprendizaje, se encuentran clasificadores que proveen un grado severamente desbalanceado de acierto, con efectividades de casi el 100 % para la clase mayoritaria pero efectividades entre $0\,\%$ y $10\,\%$, aún cuando el hecho de saber si el paciente tiene cáncer es de tanta importancia.

Muchos otros conjuntos de datos en aplicaciones reales implican bases de datos con estas características, con aplicaciones tales como la identificación de defectos de software, predicción de desastres naturales, detección de focos epilépticos en imágenes de resonancias magnéticas, transacciones de tarjetas de crédito fraudulentas y fraude de telecomunicaciones (Japkowicz and Stephen, 2002). Como se mencionó anteriormente, los algoritmos de aprendizaje de máquina con los que se suele abordar aplicaciones de este tipo, tienden a ignorar la clase minoritaria, que en estos casos es la de mayor relevancia y clasificar erróneamente eventos de esta naturaleza puede resultar en costos elevados.

Métodos de re-muestreo

Dada la gran importancia del problema del desbalance de clases en diversos campos de aplicación, distintos algoritmos y métodos han sido propuestos en la última década para abordar el problema de clasificación de datos desbalanceados. En este dominio, se requiere un clasificador que proporcione una alta precisión para la clase minoritaria

pero sin poner en peligro la precisión de la clase mayoritaria. En tal sentido, se han propuesto tres estrategias básicas: (I) *métodos de remuestreo*, los cuales son técnicas de preproceso que intentan equilibrar las distribuciones al considerar las proporciones representativas de los ejemplos de clase en la distribución, (II) *métodos de aprendizaje costo-sensitivos*, los cuales consideran los costos asociados con la clasificación errónea de las muestras (He and Garcia, 2009) y (III) *métodos de ensamble* que consisten en la combinación de dos o más clasificadores.

Las técnicas de remuestreo se utilizan para equilibrar el espacio muestral para un conjunto de datos desequilibrado con el fin de aliviar el efecto de la distribución sesgada de clase en el proceso de aprendizaje. Los métodos de remuestreo son más versátiles porque son independientes del clasificador seleccionado (López et al., 2013). Las técnicas de remuestreo se dividen en tres grupos dependiendo del método utilizado para equilibrar la distribución de clases: Métodos de sub-muestreo, métodos de sobre-muestreo y métodos híbridos.

Se define también $\boldsymbol{X}_+ \in \mathbb{R}^{N_+ \times P}$ como los datos pertenecientes a la clase mayoritaria $z = +1$ y $\boldsymbol{X}_- \in \mathbb{R}^{N_- \times P}$ como las muestras pertenecientes a la clase minoritaria $z = -1$ con $N_+ \gg N_-$. El uso de métodos de remuestreo en aplicaciones de aprendizaje desbalanceado consiste en la modificación del conjunto de datos por algunos mecanismos con el fin de proporcionar una distribución equilibrada (Galar et al., 2012), lo que significa que $N_- \equiv N_+$ después de la operación.

SMOTE

Una de las técnicas de sobremuestreo más populares en la actualidad es la técnica de creación de muestras sintéticas de la clase minoritaria SMOTE (*Synthetic Minority Oversampling Technique*). SMOTE es un método que ha mostrado tener éxito en diversas aplicaciones que involucran bases de datos desbalanceadas (Chawla et al., 2002). El algoritmo SMOTE crea datos artificales entre muestras de la clase minoritaria. Específicamente, para el subconjunto $X_- \in X$, considere los k-vecinos más cercanos de cada muestra $\boldsymbol{x_i} \in X_-$, para algún entero k; los k-vecinos más cercanos son definidos como las k muestras de X_- cuya distancia euclidea entre ellos y la muestra $\boldsymbol{x_i}$ bajo consideración presenta las menores magnitudes. Para crear una muestra sintética, se selecciona aleatoriamente uno de los k-vecinos, luego se multiplica el correspondiente vector de diferencia por un número aleatorio entre el rango $[0, 1]$, y finalmente, se suma el anterior resultado al vector $\boldsymbol{x_i}$

$$\boldsymbol{X}_{new} = \boldsymbol{x_i} + (\hat{\boldsymbol{x}}_{\boldsymbol{i}} - \boldsymbol{x_i}) \times \delta, \tag{3.1}$$

donde $\hat{\boldsymbol{x}}_{\boldsymbol{i}}$ es uno de los k-vecinos más cercanos de $\boldsymbol{x_i}$ y $\delta \in [0, 1]$ es un número aleatorio.

3.1.3. Análisis de relevancia

El objetivo principal en un sistema de reconocimiento automático de patrones consiste en separar en k diferentes clases las observaciones o datos de entrada, medidas mediante variables denominadas características (Wang and Paliwal, 2003). El número de variables empleadas para medir las observaciones se conoce como la dimensión del espacio de características. Cuando la dimensión del espacio de características es alta, existe un gran interés de reducir esta dimensión (Brunzell and Eriksson, 2000), porque uno de los problemas de tener conjuntos de alta dimensión es que, en muchos casos, no todas las variables medidas son importantes para la comprensión del fenómeno subyacente en análisis (Fodor, 2002), es decir, existen unas variables que son relevantes para la identificación del patrón y otras que no lo son.

Las razones para reducir la dimensión del espacio de características son: facilitar el análisis de los datos, mejorar el desempeño de clasificadores a través de una representación más consistente de los datos, tanto para tener una mayor precisión como para disminuir el costo computacional, remover la redundancia o información irrelevante, hacer clara cualquier estructura oculta, obtener una representación gráfica más entendible, entre otras (Carreira-Perpinán, 1997).

Cuando se habla de reducción de la dimensión, es posible distinguir entre 2 tipos de técnicas; de *selección* y de *extracción*.

Selección de características

La selección de características es el proceso de seleccionar un subconjunto de características relevantes para usar en la construcción del modelo. Las técnicas de selección son empleadas para encontrar características que son ya sea redundantes o irrelevantes, y pueden ser removidas sin pérdida de información.

Las técnicas de selección pueden ser distinguidas de las de extracción. La extracción crea nuevas características combinando las características originales, mientras la selección de características retorna un subconjunto de estas.

Un método muy utilizado para hacer análisis de relevancia y determinar las características más relevantes en una aplicación de reconocimiento de patrones, es la técnica conocida como RELIEF-F.

El algoritmo RELIEF asume que una característica es fuertemente relevante si esta permite distinguir fácilmente entre dos instancias de diferentes clases, y basándose en esta lógica define el peso para cada característica. Sin embargo, RELIEF en su versión original limita su campo de aplicación a problemas en donde solo se tienn dos clases y por lo tanto para la asignación de los pesos emplea solamente un vecino más cercano de diferente clase, debido a que solo trabaja con una clase opuesta. Por lo anterior, surgió la necesidad de ampliar el campo de aplicación del algoritmo a una nueva versión de RELIEF denominada RELIEF-F. En RELIEF-F se generaliza el comportamiento del algoritmo original para problemas donde se cuenta con más de dos clases, como en el caso del presente proyecto. En esta nueva versión del algoritmo, se busca un vecino más

cercano por cada clase opuesta. Con los vecinos seleccionados se evalúa la relevancia para cada característica y luego se actualiza su valoración acumulada en un vector de pesos.

La metodología de selección consiste básicamente en ordenar todas las características en orden descendente de acuerdo a su peso de relevancia, y se inicia un proceso heurístico para determinar el número mínimo de estas para obtener el más alto resultado posible. En esta metodología se escoge la característica de mayor peso de relevancia y se clasifica únicamente con esta. Seguidamente, se juntan la primera con la segunda y se repite el proceso de clasificación. Este método termina cuando un mayor número de características no supera los resultados de clasificación de un conjunto con una menor cantidad.

3.2. Estado de la cuestión

Para la clasificación de grupos de MCC, donde el objetivo es discriminar entre masas benignas y malignas, se han presentado principalmente dos alternativas. La primera, está basada en el entrenamiento de un clasificador a partir de un conjunto de imágenes, que previamente han sido etiquetadas y segmentadas por un especialista. Estas imágenes etiquetadas, son caracterizadas por descriptores de forma, apariencia, intensidad o textura y finalmente, son empleadas para sintonizar los parámetros de un clasificador. Por esta primera alternativa, Rouhi et al. (2015) propone extraer características a partir de los factores de forma, los momentos de zernike, histogramas de color y características Haralick. En una etapa siguiente, realiza una reducción de dimensión a partir de algoritmos genéticos, donde se selecciona un subconjunto de todas las características inicialmente calculadas.

Finalmente, las imágenes son empleadas para calcular los pesos de una red neuronal, que busca clasificar entre imágenes con masas benignas y malignas. Un planteamiento similar, propone Xie et al. (2016) que emplea características de textura e intensidad en áreas específicas de las masas etiquetadas, así como también características de forma en la región segmentada por el especialista. Empleando este conjunto de características, una máquina de aprendizaje extremo clasifica masas sin etiquetar en malignas o benignas.

La segunda alternativa, está basada en los motores de búsqueda de imagen por contenido (content-based image retrieval, CBIR). Esta alternativa busca ofrecer decisiones en el contexto de encontrar imágenes que pertenecen a una clase particular de cierta base de datos (masas malignas o benignas). Resultando en que, al radiólogo, se le está proporcionando una ayuda visual, con la que puede justificar los resultados obtenidos. Consecuentemente, la confianza al introducir sistemas CAD para la toma de decisiones aumenta. Wang et al. (2012) incorporó un clasificador k-vecinos-cercanos (k-Nearest- Neighbors, k-NN) a un sistema CAD empleando un esquema CBIR, logrando un área bajo la curva característica operativa del receptor (Receiver operating characteristic, ROC) de 0.9 ± 0.06 (Índice Az) al diferenciar 1800 imágenes de falsos positivos (300 masas benignas y 1500 con parénquima normal) (Tsochatzidis

et al., 2017). Entre las características empleadas para caracterizar todo el seno, se destacan los descriptores de masa (estadísticas longitud radial y factores de forma), así como también descriptores de intensidad (histogramas de intensidad y fluctuaciones locales de intensidad en pixeles). Wei et al. (2012) propuso recientemente un esquema CBIR en el que se pretende detectar masas (masas contra no masas) en imágenes mamográficas, basado en descriptores locales de Transformación de Característica Invariantes a la Escala (Scale-Invariant Feature Transform, SIFT) y el método de bolsa de palabras visuales. Aunque este estudio no tenía como objetivo ayudar a realizar un diagnóstico, se encontró que el trabajo desarrollado era eficiente, y podía ser fácilmente escalado.

Finalmente, para segmentar MCC en imágenes mamográficas, se han empleado diferentes modelos de substracción de fondo, que permiten discriminar entre elementos del fondo, como tejidos mamarios y elementos de primer plano, en este caso grupos de MCC. En (González et al., 2017), se presenta una alternativa para la sustracción de fondo empleando aprendizaje supervisado con procesos Gaussianos (Gaussian Processes, GP). En dicho trabajo, a partir de la clasificación de regiones en la imagen, se obtiene una imagen binaria donde las regiones asociadas a la estructura anatómica objetivo se etiquetan con "1" y las demás regiones se etiquetan con "0". Por último, para recuperar la forma de la estructura objetivo se realizan operaciones morfológicas que delimitan su contorno. Por otra parte, en Ciecholewski (2017), se realizan transformaciones morfológicas de imagen, con el objetivo de detectar microcalcificaciones, para luego obtener su forma a partir de la segmentación Watershed (Tsochatzidis et al., 2017).

3.3. Métodos

A continuación se describen los métodos y la metodología general que se siguió para clasificar imágenes de mamografías en sus respectivas escalas de BI-RADS. De manera general el procedimiento consiste en tomar los ROI de las lesiones previamente localizadas y caracterizadas como datos de entrada. Como datos de salida se encuentra la etiqueta de la mamografía, es decir su categoría, bien sea, benigna o maligna o la escala BI-RADS específica asociada a cada una de ellas. El objetivo es entrenar un modelo que aprenda a reconocer cada una de estas etiquetas a partir del procesamiento de imágenes con diagnósticos confirmados.

3.3.1. Clasificación de mamografías

Como se mencionó anteriormente, en la tarea de clasificación automática de mamografías para la detección de cáncer de mama, algunas clases se presentan con menor frecuencia que las otras, es decir, muchas de las mamografías pertenecen a una misma escala BI-RADS, mientras que muy pocas imágenes son etiquetadas en algunas escalas BI-RADS específicas. Este es precisamente el caso de algunas de las microcalcificaciones catalogadas en la BI-RADS-4, las cuales ocurren con mucha

menor frecuencia que las demás. Sin embargo, no es solo las mamografías en esta escala las que se encuentran sub-representadas. Generalmente, las bases de datos de mamografías suelen presentar una distribución no uniforme entre sus clases, ya que la ocurrencia de cada uno de los tipos de anormalidad es completamente diferente.

Debido a la poca cantidad de muestras pertenecientes a cada una de las clases, y para evitar la necesidad de eliminar muestras con información relevante, la mejor opción sin duda es hacer uso de los métodos de sobremuestreo. Estos métodos consisten en la creación de nuevas muestras de clase minoritaria. Dos métodos ampliamente utilizados para crear las muestras minoritarias sintéticas son duplicando al azar las muestras minoritarias y **SMOTE** (*Synthetic Minority Oversampling Technique*)

En este caso particular se usará la técnica de creación de muestras sintéticas de la clase minoritaria conocida como SMOTE con el objetivo de balancear el número de muestras entre cada una de las clases de la base de datos.

Para la clasificación de las bases de datos de esta investigación, se usaron tres tipos de clasificadores: En la etapa de clasificación se probaron 3 algoritmos diferentes y se comparó el rendimiento de cada uno de estos. Los algoritmos utilizados son las Máquinas de Vectores de Soporte (SVM), bosques aleatorios (Random Forest) y el método conocido como Adaboost. Estos tres métodos son clasificadores ampliamente usados para el manejo de conjuntos de datos que exhiben una distribución desbalanceada de sus clases.

También, se evalúa el desempeño en la clasificación como la cantidad de aciertos de cada clasificador usando un esquema de validación cruzada de 5 *folds*. Esto es, la base de datos se divide aleatoriamente en 5 grupos. Para cada grupo, los algoritmos son entrenados con las muestras de los 4 grupos restantes y luego se prueba el clasificador sobre el conjunto actual.

3.4. Resultados

Los resultados de clasificación se encuentran en las tablas 3.1 hasta la 3.3. En estas tablas se presentan las matrices de confusión de cada uno de los tres clasificadores, representadas en proporción de aciertos para cada una de las clases. Las clases 0, 2, 3, 4 y 5 corresponden a cada una de las escalas BI-RADS de la base de datos de entrenamiento.

De los resultados se puede observar que cuanto más alto los valores de la diagonal principal de la matriz de confusión, más efectivo es el clasificador para detectar cada una de las clases, y mientras más altos los valores fuera de la diagonal, indica un mayor número de falsos positivos o desaciertos del clasificador. De acuerdo a esto, la SVM parece ser el algoritmo que realiza mejor la tarea de clasificación automática de mamografías en la escala BI-RADS a partir de las características previamente extraídas.

		Predicción				
		0	**2**	**3**	**4**	**5**
Valor verdadero	**0**	100	0	0	0	0
	2	0	83.5	11.3	0.9	4.3
	3	0	9.6	89.6	0.9	0
	4	0	3.5	0	96.5	0
	5	0	3.5	0.9	0	95.7

Cuadro 3.1:
Resultados clasificación SVM.

		Predicción				
		0	**2**	**3**	**4**	**5**
Valor verdadero	**0**	100	0	0	0	0
	2	0	78.3	10.4	4.3	7
	3	0	12.2	86.1	1.7	0
	4	0	6.1	0.9	92.2	0.9
	5	0	3.5	1.7	0	94.8

Cuadro 3.2:
Resultados clasificación Bosques aleatorios.

		Predicción				
		0	**2**	**3**	**4**	**5**
Valor verdadero	**0**	100	0	0	0	0
	2	0.9	62.6	18.3	12.2	6.1
	3	0.9	15.7	77.4	5.2	0.9
	4	0	13.9	7	76.5	2.6
	5	0.9	2.6	3.5	2.6	90.4

Cuadro 3.3:
Resultados clasificación Adaboost.

3.5. Conclusiones

En conclusión, en este proyecto se desarrolló un sistema de clasificación de la escala BI-RADS a partir del análisis, caracterización y procesamiento de imágenes de mamografías digitales. El sistema se encarga de crear muestras sintéticas de imágenes mamográficas de cada una de las escala BI-RADS de tal manera que cada una tenga la misma cantidad de muestras a través del método de remuestreo SMOTE. Adicionalmente, el sistema determina cuáles son las variables que tienen mayor capacidad predictiva y se eliminaron aquellas irrelevantes. Entre los tres clasificadores probados, la SVM multiclase demostró ser el más eficiente, mostrando precisiones en cada una de las clases superiores al 80 %. Específicamente, en la escala BI-RADS

2 que suele ser una de las más complicadas de detectar es el modelo que mejor la predice. Finalmente, este sistema ya entrenado, permite que una nueva imagen pueda ser procesada y clasificada, dependiendo de la decisión de este modelo, lo cual a su vez es un importante insumo para el sistema de recomendación.

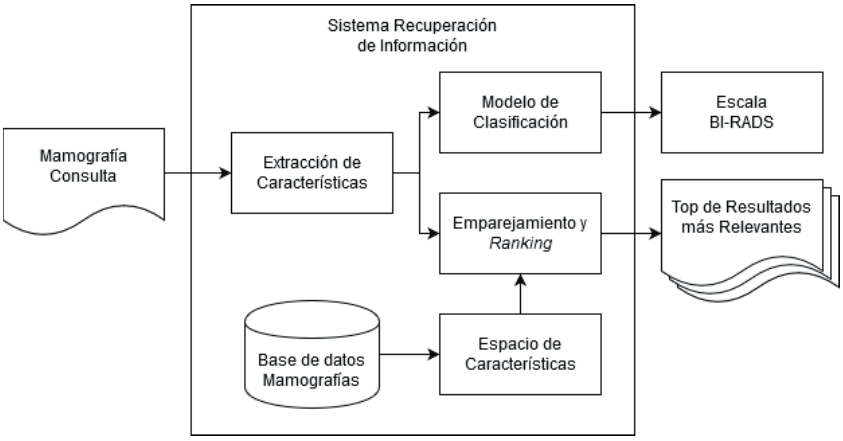
4 CAPÍTULO CUATRO

4. Sistema de recomendación para la recuperación de información

En este capítulo se describen las técnicas empleadas en el sistema de recuperación de información. Así mismo, se presenta el estado actual de las investigaciones más relacionadas en esta área y se presenta una revisión de los resultados más relevantes encontrados en otros proyectos de investigación. En la sección 4.1 se presenta el marco teórico de las técnicas empleadas en los sistemas de recomendación basado en características. En la sección 4.2 se presentará el estado de la cuestión relacionado con las investigaciones más relevantes que se han realizado en la misma línea de este trabajo. En la sección 4.3 se habla de los métodos usados para llevar a cabo la investigación; y finalmente, en la sección 4.4 se presentan los resultados obtenidos.

El diagrama de flujo presentado en la figura 4.1 muestra las diferentes etapas del sistema de recuperación de información.

Figura 4.1:
Diagrama de flujo del Sistema de Recuperación de Información



De forma general, el sistema de recuperación de información está compuesto por tres módulos: el primer módulo (**extracción de características**) se encarga de

obtener atributos de tipo morfológico y de textura de cada ROI para caracterizar las microcalcificaciones; el segundo módulo (**emparejamiento y *ranking***) determina las imágenes a recuperar en función de una medida de similitud y las clasifica en orden de relevancia; y el tercer módulo (**clasificación**) se encarga de entregar una etiqueta en función de la escala BI-RADS determinada de manera automática en la mamografía de consulta.

4.1. Marco teórico

La recuperación de imágenes basada en contenido (CBIR - *Content Based Information Retrieval*) ha sido empleada frecuentemente en el manejo de imágenes en distintas aplicaciones, sin embargo, se han desarrollado pocos sistemas de recuperación basados en contenido específicamente para imágenes médicas, como el sistema ASSERT (Chen and Zwiggelaar, 2010; Shyu et al., 1999) y el proyecto NHANES II (Antani et al., 2004,0). Estos sistemas generalmente se construyen en institutos de investigación y continúan siendo mejorados, desarrollados y evaluados con el tiempo.

Qi and Snyder (1999) destacaron los inconvenientes existentes en la recuperación textual tradicional y abordan la importancia de la recuperación de imágenes archivadas. Como resultado, se implementó un sistema CBIR que hace uso de mamografias digitales, donde la recuperación se hace en función de la forma del objeto, tamaño y brillo del histograma de las masas. Sin embargo, este estudio no proporcionó más discusiones sobre la arquitectura del sistema propuesto y los descriptores mamográficos en detalle.

Kinoshita et al. (2007) desarrolló un sistema de recuperación de información basado en mamografías, en dicho sistema se utilizaron características como la densidad mamaria, patrones basados en el tamaño anatómico y la forma de la región mamaria, características estructurales y distribuciones de densidad del tejido fibro-glandular, así como características menos exploradas basadas en el dominio de Radon y las medidas granulométricas. Para realzar la operación de recuperación utilizaron los mapas auto-organizados de Kohonen. Sin embrago, aunque muchas de las características se extrajeron para representar las propiedades visuales en mamografías, el sistema propuesto mide todas esas características en conjunto sin tener en cuenta su relevancia. Tal estructura de medida es probable que resultara en una gran desventaja ya que una propiedad particular representada por un mayor número de características dominará otras propiedades descritas por solo unas pocas características.

En otro estudio de Alto et al. (2005) se investigó el uso de la forma, la nitidez de los bordes y las características de textura para recuperar imágenes con masas similares. Las características fueron evaluadas con el fin de clasificar las masas como benignas o malignas empleando análisis discriminante lineal, regresión logística y un clasificador basado en la distancia de Mahalanobis. El análisis discriminante lineal entregó la mayor sensibilidad (del $100\,\%$) y especificidad (del $97\,\%$).

Muramatsu et al. (2007) investigó un medida de similitud psicofísica para comparar

imágenes similares a las de masas desconocidas en mamografías. En este estudio se les pidió a cinco radiólogos evaluar 60 pares de masas basados en calificaciones subjetivas de similitud, que fueran marcadas en una escala continua y cuantificada entre 0 y 1, de manera paralela, se calcularon las distancias euclidianas entre las características de las masas desconocidas. Posteriormente se obtuvo la medida de similitud psicofísica entrenando una red neuronal artificial a partir de la relación entre los índices de similitud subjetiva de los radiólogos y las correspondientes distancias euclidianas de las masas en el espacio de características. El principal inconveniente es que se necesita una gran cantidad de datos para entrenar una red neuronal artificial en el proceso de aprendizaje automático. Además, existen otros inconvenientes comunes en este tipo de estudios, aunque esos estudios proponen métodos sobre lesiones específicas, como el aprendizaje de retroalimentación de relevancia para recuperar masas similares, la metodología para el diseño un sistema CBIR completo no se ha propuesto, lo que implica que el preprocesamiento, detección y segmentación de lesiones, extracción de características no se incluyen en estos tipos estudios.

4.2. Estado de la cuestión

Una de las cuestiones más importantes en el desarrollo de un sistema de recuperación de información en mamografías es que las características extraídas deben ser lo suficientemente discriminatorias como para representar diferentes tipos de características patológicas. Dado que la presencia de microcalcificaciones se ha considerado como un indicador crucial para el diagnóstico de los hallazgos en tejidos mamarios, se han realizado muchas investigaciones para desarrollar métodos confiables para la identificación precisa de las mismas. Para detectar microcalcificaciones agrupadas, Arodź et al. (2006) filtran la mamografía con un filtro que es sensible a la forma de la microcalcificación y mejoran el contraste mediante el uso de un algoritmo de nitidez basado en wavelets.

El-Naqa et al. (2004) utilizan un método de máquinas de vectores de soporte (SVM) para detectar grupos de microcalcificaciones. Las características que utilizaron incluyen área de sección transversal, compacidad, excentricidad, densidad, dispersión, solidez, momento invariante, firma de momento y descriptor de Fourier normalizado. Peng et al. (2006) incorporan un mecanismo de descubrimiento de información en un algoritmo genético para detectar microclacificaciones. Nueve características están diseñadas para describir puntos brillantes en términos de su forma y textura.

Linguraru et al. (2006) desarrollaron un método para la detección de microcalcificaciones basado en un modelo adaptativo de detección de contraste, una novedad importante de ese trabajo es que su algoritmo puede estimar automáticamente los valores de los parámetros a partir de la imagen. Como la distribución espacial y la forma de las microcalcificaciones tienen un impacto significativo en la práctica médica, Bocchi and Nori (2007) utilizaron la transformación de Radón para desarrollar un conjunto de características, evaluando así la morfología de las manchas calcificadas. Además, Nakayama et al. (2006) y Retico et al. (2006) desarrollaron un

esquema de detección asistido por computadora para la identificación del grupo de microcalcificaciones.

Además de la extracción de características mamográficas, el desarrollo de un sistema de recuperación de imágenes debe tener en cuenta los factores humanos; entre estos, la percepción subjetiva es uno de los problemas más desafiantes. Una imagen es una representación simbólica; las personas interpretan una imagen y asocian la semántica con ella en función de sus percepciones subjetivas, lo que implica el conocimiento del usuario, sus antecedentes culturales, sus sentimientos personales, etc. (Jaimes and Dimitrova, 2006). Para compensar la vaguedad de las percepciones humanas subjetivas, los sistemas de recuperación de imágenes deben poder interactuar con los usuarios y descubrir las necesidades de información del usuario actual, por medio de la retroalimentación de la relevancia de los resultados. La retroalimentación de relevancia hace referencia a cómo el sistema de recuperación puede darse cuenta de la necesidad de información del usuario. Esta realimentación se realiza analizando la relevancia que han tenidos los resultados para el usuario y conectando la necesidad de información del usuario con características de bajo nivel para mejorar los resultados de búsqueda.

Tres enfoques de retroalimentación de relevancia son: el movimiento del punto de consulta, la revaloración y la clasificación.

El enfoque de movimiento del punto de consulta supone que existe al menos una imagen que transmite completamente las intenciones del usuario, y su concepto de alto nivel se ha modelado en el espacio de características de bajo nivel (Rui, 1998). El enfoque de movimiento del punto de consulta implica mover el punto de la consulta hacia la región del espacio de características que contiene la imagen ideal; como cada imagen está representada por un vector de características de n dimensiones, el vector de características puede considerarse como un punto en un espacio de n dimensiones.

La idea del enfoque de revaloración es ajustar los pesos asignados a cada característica o modificar la medida de similitud utilizada (Rui and Huang, 1999), es decir, dar a las características importantes pesos más grandes y a las características menos importantes pesos más pequeños.

Dado que las imágenes con características similares se encuentran juntas en el espacio de características, las imágenes se pueden clasificar en diferentes categorías según sus características. El enfoque de clasificación implica el uso de información de retroalimentación para clasificar todo el conjunto de imágenes (Gondra et al., 2004). Los que pertenecen a una clase dada pueden considerarse imágenes similares.

4.3. Métodos

A continuación, se describen los métodos y la metodología general que se siguió para la implementación de un sistema de recuperación basado en imágenes mamográficas.

De forma general el sistema está compuesto por dos etapas: la primera de ellas es la etapa de extracción de características (ver sección 4.3.2), la cual consiste en representar a cada ROI como un vector de atributos que describan las microcalcificaciones

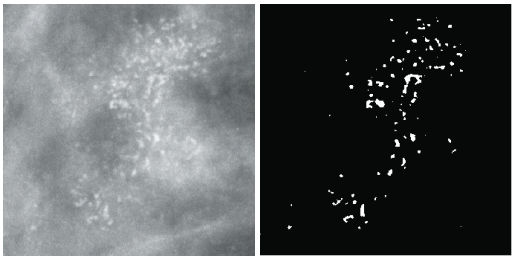
contenidas en la imagen. La segunda es la etapa de emparejamiento (ver sección 4.3.3), en donde se calcula la similitud entre una ROI de entrada y el resto del conjunto de imágenes en el repositorio del sistema.

El objetivo de esta metodología es implementar una interfaz que permita a un usuario seleccionar una ROI afectada con microcalcificaciones y que el sistema retorne un determinado subconjunto de imágenes similares a la imagen consultada por el usuario (esto se explica más en detalle en la sección 4.3.4).

4.3.1. Preprocesamiento de la ROI

El módulo de preprocesamiento de la ROI se basa en 3 etapas, las cuales se ejecutan de forma semi-automática. La primera consiste en remover el fondo de la ROI; para esto se aplica una transformación morfológica de apertura a la ROI original, usando como elemento estructurante un disco de radio variable con el fin de determinar la información correspondiente al fondo de la imagen (Zhao, 1993). La segunda etapa consiste en la reconstrucción de la ROI a partir de los coeficientes de aproximación obtenidos de una descomposición Wavelet; el objetivo es obtener una ROI donde el contraste de las microcalcificaciones es realzado por el incremento en la intensidad de las regiones donde estas se encuentran, filtrando todo tipo de estructuras y tejidos irrelevantes (Alam et al., 2018; G., 2006). En la tercera etapa se binariza la ROI mediante una técnica de umbralización manual, así se obtiene el número de objetos (microcalcificaciones) que contiene la ROI, además se usa un filtro de tamaño 3×3 para eliminar el ruido sobrante.

Figura 4.2:
Preprocesamiento del ROI.



(a) ROI original (b) ROI preprocesado

4.3.2. Extracción de características

Los descriptores visuales son el primer paso para encontrar la conexión entre los píxeles contenidos en una imagen. Estos descriptores se dividen en dos grupos principales: a) Descriptores de información general, los cuales contienen descriptores de bajo nivel que representan el color, la forma, las regiones, las texturas, el

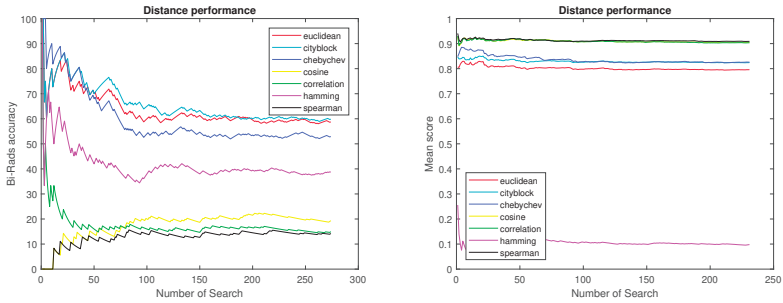
movimiento, etc.; y b) los descriptores de información específica del entorno, los cuales proporcionan información sobre objetos y eventos en la escena. Los descriptores pueden organizarse en vectores de características (VC) los cuales contienen toda la información extraída de la imagen (Estrela and Herrmann, 2016).

En la etapa de extracción de características de este trabajo se usaron dos extractores. El primero de ellos está encargado de extraer características (21 en total) de tipo morfológico, tales como: área, intensidad promedio, excentricidad, el ratio entre eje mayor y eje menor, el ratio entre los centroides ponderados, en adición a esto todos los valores medios, desviaciones estándar, coeficientes de variación, valores máximos y mínimos de las características mencionadas. Cada una de estas características es tomada de cada microcalcificación de forma individual (Chan et al., 1998). El segundo extractor se encargó de extraer las características relacionadas con la textura de la imagen (22 en total), éstas fueron: uniformidad, energía, segundo momento angular, entropía, disimilaridad, contraste, inercia, diferencia inversa, correlación, homogeneidad, autocorrelación, sombra y prominencia del cluster, probabilidad máxima, suma de cuadrados, promedios, varianza y entropía, información medida de la correlación y autocorrelación, coeficiente de máxima correlación y momento normalizado (Clausi, 2002).

4.3.3. Emparejamiento

En el proceso de emparejamiento, el vector de características de la imagen consulta (la imagen que el usuario ingresa y de la cual el sistema CBIR usa como referencia para retornar imágenes similares) se compara con los VC de las otras imágenes almacenadas en la base de datos. Esto implica el cálculo de alguna medida de similitud entre VC para posteriormente ordenar en un *ranking* las imágenes acorde al nivel de similitud (Estrela and Herrmann, 2016). En este trabajo se calcula el nivel de similitud empleando las distancias Euclidiana, Cityblock, Chebyshev, Coseno, Correlación, Hamming y Spearman. Posteriormente se analiza su rendimiento, en términos del nivel de acierto en relación al BI-RADS, esto es que la imagen consulta sea emparejada con otra imagen perteneciente a la misma clase de BI-RADS, como se puede observar en la figura 4.3 (a). Por otro lado, se mide el nivel de similitud promedio con imágenes que se consideran primera opción en el emparejamiento, es decir imágenes cuyo VC se ubica mas próximo al VC de la imagen consulta en el espacio de características como se indica en la figura 4.3 (b). En este experimento se usó una base de datos que contenía 107 ROI extraídas de la base de datos UTP, las cuales ya han sido categorizadas previamente por expertos según el estándar BI-RADS. Los resultados muestran que la distancia que mejor empareja es la Euclidiana, tal como puede verse en la figura 4.3.

Figura 4.3:
Rendimiento de cada medida de similitud.



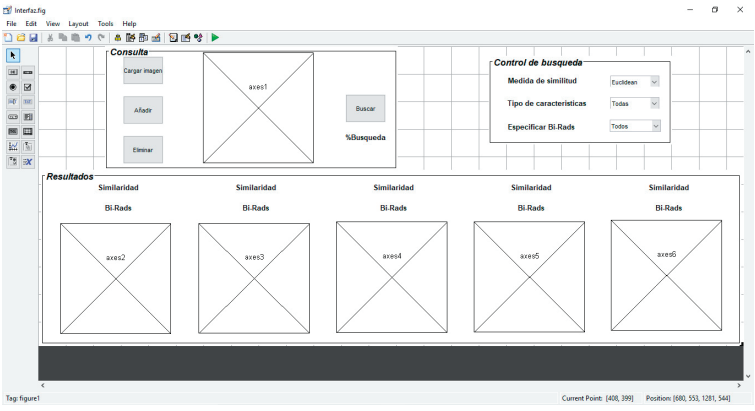
(a) Rendimiento en función del nivel de acierto en el BI-RADS.

(b) Rendimiento en función de la similitud promedio.

4.3.4. Interfaz gráfica

Luego de que las etapas de caracterización y emparejamiento se implementaron en MATLAB, estas se acoplaron para conformar el sistema de CBIR. La interfaz se diseñó con GUI Layout Toolbox (MATLAB). Esta aplicación proporciona herramientas para crear interfaces gráficas de usuario sofisticadas, suministrando diversos componentes (botones, sliders, paneles, menús deslizantes, visualizador de imágenes, etc.) que pueden ser utilizados en combinación para producir prácticamente cualquier diseño de interfaz de usuario. A continuación se presenta, en la figura 4.4, la interfaz diseñada para el sistema CBIR.

Figura 4.4:
Diseño de la interfaz gráfica del sistema CBIR.



En la figura 4.4 se muestra el diseño de la interfaz gráfica, el cual está compuesto por tres paneles principales. El primer panel es el de consulta, éste está diseñado para que el usuario pueda especificar la imagen consulta usando el botón **Cargar imagen**, este panel también permite visualizar la imagen consulta además de añadir la imagen consulta a la base de datos relacionada al sistema CBIR o eliminarla si esta ya hacía parte de la misma. Adicionalmente, en este panel se encuentra el botón de **Buscar** el cual le permite al usuario inicializar una nueva búsqueda una vez haya seleccionado la imagen consulta.

El segundo panel es el de resultados, en éste se podrán visualizar los resultados arrojados por el sistema CBIR una vez que el usuario haya realizado una consulta. Este panel, además de permitir visualizar las 5 imágenes mejor posicionadas en el *ranking* retornadas por el sistema, también indica el nivel de similitud (*score*) calculado por el sistema para cada una de las imágenes retornadas y su respectiva categoría BI-RADS. Esto se hace con el objetivo de que el usuario pueda tener información de tipo visual, presentada en las imágenes y cuantificada en términos de similitud; además de facilitar la categorización del BI-RADS al poder tomar como referencia las categorías de las imágenes retornadas.

El tercer panel es el de **Control de búsqueda**, aquí el usuario podrá especificar la medida de similitud con la cual el sistema realizará el emparejamiento. Las opciones a elegir son las mismas con las que se realizó el experimento de la sección 4.3.3, sin embargo, por defecto, el sistema utilizará la métrica Euclidiana ya que esta conduce a mejores resultados. Por otro lado, en este panel también se puede especificar el tipo de características que conforman el VC que representa cada imagen; entre las opciones se puede seleccionar las características de tipo morfológico o las características de textura (GLCM), aunque el sistema por defecto usará ambas. Adicionalmente, el usuario también podrá determinar el "sector" de la base de datos en el cual el sistema emplee la búsqueda, esto se logra mediante la opción de especificación de BI-RADS donde el usuario puede especificar una categoría y el sistema retornará solo imágenes pertenecientes a la misma.

4.4. Resultados

Figura 4.5:
Búsqueda de un cluster de microcalcificaciones.

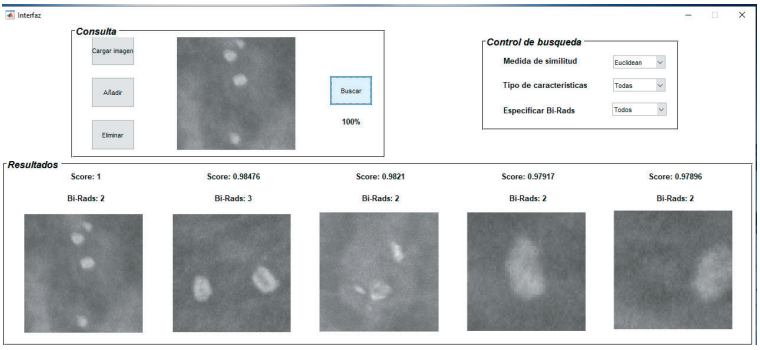
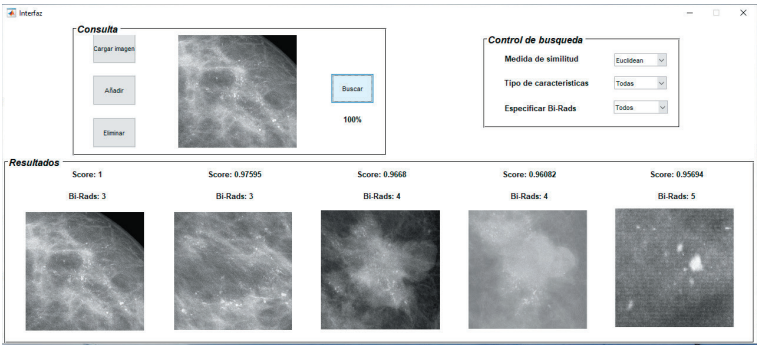


Figura 4.6:
Búsqueda de un cluster complejo de microcalcificaciones



4.5. Conclusión

En este capítulo se presentó el sistema de recuperación de información basado en imágenes. El sistema propuesto fue desarrollado para realizar consultas a partir de una imagen mamográfica en la cual existen clusters de microcalcificaciones, y retornar un conjunto de imágenes mamográficas con clusters de microcalcificaciones similares. Para evaluar el nivel de similitud entre dos clusters de microcalcificaciones se extrajeron 43 características en las que se incluyen de tipo morfológico y de textura, posteriormente se calculó el nivel de similitud entre el vector de características de la

imagen consulta y el resto de representaciones de la base de datos usando la métrica euclidiana.

Los resultados demostraron que el sistema de recuperación de información retorna imágenes de clusters visualmente similares a la imagen consulta, además usando la métrica euclidiana se obtiene un nivel de similitud promedio 80 % y un nivel de coincidencia del 60 % en la identificación del BI-RADS del cluster de microcalcificaciones.

5. Conclusiones

En este libro se presentaron los resultados de investigación del proyecto titulado "Prototipo de un Sistema de Recuperación de Información por Contenido Orientado a la Localización y Clasificación de Grupos de Microcalcificaciones en Mamografías - PROTOCAM". El objetivo principal de este proyecto era el desarrollo de una metodología que pudiera facilitarle al especialista la consulta de diagnósticos de relevancia similares a un caso concreto. Para lograr esto se implementaron exitosamente algunas de las técnicas más conocidas y utilizadas en el estado del arte que se evaluaron sobre sobre dos bases de datos, una privada de la Universidad Tecnológica de Pereira y la base de datos de uso público MIAS. Este sistema constó de 3 etapas básicas, que fueron el preprocesamiento de la mamografía, la clasificación en escala BIRADS y el sistema de recuperación las cuales en conjunto conforman el sistema de recuperación de información con el fin de localizar y detectar microcalcificaciones en imágenes de mamografías.

En la etapa de preprocesamiento se ajustaron los parámetros de los filtros teniendo como referencia las imágenes de la base de datos de la UTP. En cuanto a la eliminación de artefactos y remoción del músculo pectoral los resultados indican un buen rendimiento, lo cual se debe a la estrategia utilizada basada en una búsqueda iterativa en función de la homogeneidad de la región, con el objetivo de encontrar los parámetros óptimos en la segmentación de cada mamografía. Por otro lado, el efecto de filtrar y eliminar las regiones no deseadas de la mamografía en conjunto con las propiedades de abstracción y reconocimiento de patrones de las CNN le permiten al sistema de detección de potenciales microcalcificaciones obtener una identificación del 83.3 % de los ROIs que contenían microcalcificaciones. Estos resultados le permiten al módulo entregar recomendaciones más precisas al sistema, pues permite realizar mejores caracterizaciones de las mamografías nuevas para su posterior clasificación y comparación con la base de datos.

En cuanto al módulo de clasificación, uno de los desafíos mas grandes que se encontraron al usar técnicas de *machine learning* fue el desbalance de clases entre los tipos de escalas BIRADS de cada una de las imágenes que había en la base de datos de la UTP. El método de balanceo utilizado en conjunto con la técnica de análisis de relevancia y el clasificador entrenado arrojaron resultados muy satisfactorios, con resultados de precisión mayores al 80 % en cada una de las clases. Estos resultados en el sistema final se utilizan como un metadato para ayudar al sistema a refinar sus recomendaciones y obtener mejores resultados.

El sistema de recuperación de información que se describió en este libro, obtuvo un resultado de similitud en términos de la proximidad media a la imagen consulta de 80 % y una coincidencia de la categoría BIRADS con la imagen consulta del 60 %. Estos resultados se deben al buen rendimiento de los módulos que componen las etapas anteriores del procesamiento; al igual que a la exhaustiva búsqueda que se realiza en toda la base de datos. El algoritmo de emparejamiento basado en la métrica Euclidiana que el sistema incorpora le brindará al especialista una herramienta que aumentará la certeza de su diagnóstico, esto significará una mejor experiencia tanto para el especialista como para el paciente.

Como trabajo futuro se espera mejorar cada una de las etapas para aumentar la capacidad de discriminación y recomendación del sistema, además de la utilización de más técnicas de *machine learning* como redes neuronales para la identificación y segmentación de microcalcificaciones.

Bibliografía

Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

Nashid Alam, Arnau Oliver, Erika RE Denton, and Reyer Zwiggelaar. Automatic segmentation of microcalcification clusters. In *Annual Conference on Medical Image Understanding and Analysis*, pages 251–261. Springer, 2018.

Hilary Alto, Rangaraj M Rangayyan, and JE Leo Desautels. Content-based retrieval and analysis of mammographic masses. *Journal of Electronic Imaging*, 14(2): 023016, 2005.

Sameer Antani, DJ Lee, L Rodney Long, and George R Thoma. Evaluation of shape similarity measurement methods for spine x-ray images. *Journal of Visual Communication and Image Representation*, 15(3):285–302, 2004.

Sameer K Antani, Xiaoqian Xu, L Rodney Long, and George R Thoma. Partial shape matching for cbir of spine x-ray images. In *Storage and Retrieval Methods and Applications for Multimedia 2004*, volume 5307, pages 1–8. International Society for Optics and Photonics, 2003.

Tomasz Arodź, Marcin Kurdziel, Tadeusz J Popiela, Erik OD Sevre, and David A Yuen. Detection of clustered microcalcifications in small field digital mammography. *computer methods and programs in biomedicine*, 81(1):56–65, 2006.

National Electrical Manufacturers Association et al. Determination of image uniformity in diagnostic magnetic resonance images. *NEMA Standard Publications, MS 3*, 2005.

P Athira, K K Fasna, and Anjaly Krishnan. An Overview Of Mammogram Noise And Denoising Techniques. *International Journal of Engineering Research and General Science*, 4(2):557–563, 2016.

Derya Avci and Akif Dogantekin. An expert diagnosis system for parkinson disease based on genetic algorithm-wavelet kernel-extreme learning machine. *Parkinson's disease*, 2016, 2016.

R. Baeza-Yates and B. Ribeiro-Nieto. *Modern Information Retrieval*, volume 2. Addison Wesley, 2011.

R Baker, KD Rogers, N Shepherd, and N Stone. New relationships between breast microcalcifications and cancer. *British journal of cancer*, 103(7):1034, 2010.

TMA Basile, A Fanizzi, L Losurdo, R Bellotti, U Bottigli, R Dentamaro, V Didonna, A Fausto, R Massafra, M Moschetta, et al. Microcalcification detection in full-field digital mammograms: A fully automated computer-aided system. *Physica Medica*, 64:1–9, 2019.

Nina Bijker, M Donker, J Wesseling, GJ den Heeten, and EJ Th Rutgers. Is dcis breast cancer, and how do i treat it? *Current treatment options in oncology*, 14(1):75–87, 2013.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

L Bocchi and J Nori. Shape analysis of microcalcifications using radon transform. *Medical engineering & physics*, 29(6):691–698, 2007.

Håkan Brunzell and Jonny Eriksson. Feature reduction for classification of multidimensional data. *Pattern Recognition*, 33(10):1741–1748, 2000.

Miguel A Carreira-Perpinán. A review of dimension reduction techniques. *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, 9:1–69, 1997.

Heang-Ping Chan, Berkman Sahiner, Kwok Leung Lam, Nicholas Petrick, Mark A Helvie, Mitchell M Goodsitt, and Dorit D Adler. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. *Medical physics*, 25(10):2007–2019, 1998.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Zhili Chen and Reyer Zwiggelaar. Segmentation of the breast region with pectoral muscle removal in mammograms. *Medical Image Understanding and Analysis (MIUA)*, pages 71–76, 2010.

Zhili Chen, Harry Strange, Arnau Oliver, Erika RE Denton, Caroline Boggis, and Reyer Zwiggelaar. Topological modeling and classification of mammographic microcalcification clusters. *IEEE transactions on biomedical engineering*, 62(4): 1203–1214, 2014.

Marcin Ciecholewski. Microcalcification segmentation from mammograms: A morphological approach. *Journal of digital imaging*, 30(2):172–184, 2017.

David A Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of remote sensing*, 28(1):45–62, 2002.

Jyoti Dabass. Pectoral muscle and breast density segmentation using modified region growing and k-means clustering algorithm. In *Data Communication and Networks*, pages 331–339. Springer, 2020.

Instituto Nacional de Cancerología. Hechos y acciones, vol. 4, no. 2, 2011.

Instituto Nacional de Cancerología. Plan decenal para el control del cáncer en colombia. bogotá, colombia: Ministerio de salud y protección social, 2012.

Organización Mundial de la Salud. Nota descriptiva de la oms sobre cáncer. http://www.who.int/mediacentre/factsheets/fs297/es/, 2018. Accessed: 2018-04-13.

Organización Mundial de la Salud. Cáncer de mama: prevención y control. https://www.who.int/topics/cancer/breastcancer/es/, 2020. Accessed: 2020-05-01.

Ministerio de Salud y Protección Social. Plan decenal de salud pública 2012 - 2021. https://www.minsalud.gov.co/plandecenal/Paginas/home2013.aspx, 2012.

Ministerio de Salud y Protección Social-Colciencias. Guía de práctica clínica (gpc) para la detección temprana, tratamiento integral, seguimiento y rehabilitación del cáncer de mama, 2013.

J Dheeba and S Tamil Selvi. Classification of malignant and benign microcalcification using svm classifier. In *2011 International Conference on Emerging Trends in Electrical and Computer Technology*, pages 686–690. IEEE, 2011.

Giuseppe Di Leo, Consolatina Liguori, Antonio Pietrosanto, and Paolo Sommella. A vision system for the online quality monitoring of industrial manufacturing. *Optics and Lasers in Engineering*, 89:162–168, 2017.

Issam El-Naqa, Yongyi Yang, Miles N Wernick, Nikolas P Galatsanos, and Robert M Nishikawa. A support vector machine approach for detection of microcalcifications. *IEEE transactions on medical imaging*, 21(12):1552–1563, 2002.

Issam El-Naqa, Yongyi Yang, Nikolas P Galatsanos, Robert M Nishikawa, and Miles N Wernick. A similarity learning approach to content-based image retrieval: application to digital mammography. *IEEE transactions on Medical Imaging*, 23 (10):1233–1244, 2004.

Vania Vieira Estrela and Albany E Herrmann. Content-based image retrieval (cbir) in remote clinical diagnosis and healthcare. In *Encyclopedia of E-Health and Telemedicine*, pages 495–520. IGI Global, 2016.

Imola K Fodor. A survey of dimension reduction techniques, 2002.

Damian Alberto Alvarez G. *Deteccion de microcalcificaciones en mamografias digitales*. PhD thesis, Universidad Tecnológica de Pereira. Facultad de Ingenierías Eléctrica . . . , 2006.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2012.

M Jayesh George and S Perumal Sankar. Efficient preprocessing filters and mass segmentation techniques for mammogram images. In *2017 IEEE International Conference on Circuits and Systems (ICCS)*, pages 408–413. IEEE, 2017.

Iker Gondra, Douglas R Heisterkamp, and Jing Peng. Improving image retrieval performance by inter-query learning with one-class support vector machines. *Neural Computing & Applications*, 13(2):130–139, 2004.

Julián Gil González, Andrés M Álvarez, Andrés F Valencia, and Álvaro A Orozco. Automatic peripheral nerve segmentation in presence of multiple annotators. In *Iberoamerican Congress on Pattern Recognition*, pages 246–254. Springer, 2017.

D Surya Gowri and T Amudha. A review on mammogram image enhancement techniques for breast cancer detection. In *2014 International Conference on Intelligent Computing Applications*, pages 47–51. IEEE, 2014.

Benjamin Graham. Fractional max-pooling. *arXiv preprint arXiv:1412.6071*, 2014.

Murat Alparslan Gungor and Irfan Karagoz. The effects of the median filter with different window sizes for ultrasound image. *2016 2nd IEEE International Conference on Computer and Communications, ICCC 2016 - Proceedings*, pages 549–552, 2017. 10.1109/CompComm.2016.7924761.

Ya'nan Guo, Min Dong, Zhen Yang, Xiaoli Gao, Keju Wang, Chongfan Luo, Yide Ma, and Jiuwen Zhang. A new method of detecting micro-calcification clusters in mammograms using contourlet transform and non-linking simplified pcnn. *Computer methods and programs in biomedicine*, 130:31–45, 2016.

Robert M Haralick and Linda G Shapiro. *Computer and robot vision*, volume 1. Addison-wesley Reading, 1992.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

P Henrot, A Leroux, C Barlier, and P Génin. Breast microcalcifications: the lesions in anatomical pathology. *Diagnostic and interventional imaging*, 95(2):141–152, 2014.

Marc J Homer and Edward A Sickles. *Mammographic interpretation: a practical approach*, volume 224. McGraw-Hill New York, 1997.

Alejandro Jaimes and N Dimitrova. Human-centered multimedia: culture, deployment, and access. *IEEE MultiMedia*, 13(1):12–19, 2006.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

Nico Karssemeijer, Johannes D Otten, Antonius AJ Roelofs, Sander van Woudenberg, and Jan HCL Hendriks. Effect of independent multiple reading of mammograms on detection performance. In *Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment*, volume 5372, pages 82–89. International Society for Optics and Photonics, 2004.

Sérgio Koodi Kinoshita, Paulo Mazzoncini de Azevedo-Marques, Roberto Rodrigues Pereira, Jośe Antônio Heisinger Rodrigues, and Rangaraj Mandayam Rangayyan. Content-based retrieval of mammograms using visual features related to breast density patterns. *Journal of Digital Imaging*, 20(2):172–190, 2007.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Ilhame Ait Lbachir, Rachida Es-Salhi, Imane Daoudi, and Saadia Tallal. A new mammogram preprocessing method for computer-aided diagnosis systems. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 166–171. IEEE, 2017.

Xinbin Li, Hewei Gao, Zhiqiang Chen, Li Zhang, Xiaohua Zhu, Shengping Wang, and Weijun Peng. Diagnosis of breast cancer based on microcalcifications using grating-based phase contrast ct. *European radiology*, 28(9):3742–3750, 2018.

Marius George Linguraru, Kostas Marias, Ruth English, and Michael Brady. A biologically inspired algorithm for microcalcification cluster detection. *Medical image analysis*, 10(6):850–862, 2006.

Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250: 113–141, 2013.

Maher Maalouf and Theodore B Trafalis. Rare events and imbalanced datasets: an overview. *International Journal of Data Mining, Modelling and Management*, 3(4): 375–388, 2011.

Priscilla Machado, John R Eisenbrey, Barbara Cavanaugh, and Flemming Forsberg. New image processing technique for evaluating breast microcalcifications: a comparative study. *Journal of Ultrasound in Medicine*, 31(6):885–893, 2012.

Priscilla Machado, John R Eisenbrey, Barbara Cavanaugh, and Flemming Forsberg. Microcalcifications versus artifacts: initial evaluation of a new ultrasound image processing technique to identify breast microcalcifications in a screening population. *Ultrasound in medicine & biology*, 40(9):2321–2324, 2014.

Priscilla Machado, John R Eisenbrey, Maria Stanczak, Barbara C Cavanaugh, Lisa M Zorn, and Flemming Forsberg. Ultrasound detection of microcalcifications in surgical breast specimens. *Ultrasound in medicine & biology*, 44(6):1286–1290, 2018.

E Malar, A Kandaswamy, D Chakravarthy, and A Giri Dharan. A novel approach for detection and classification of mammographic microcalcifications using wavelet analysis and extreme learning machine. *Computers in biology and medicine*, 42(9): 898–905, 2012.

Claudio Marrocco, Mario Molinara, Ciro D'Elia, and Francesco Tortorella. A computer-aided detection system for clustered microcalcifications. *Artificial intelligence in medicine*, 50(1):23–32, 2010.

Elizabeth S McDonald, Anne Marie McCarthy, Susan P Weinstein, Mitchell D Schnall, and Emily F Conant. Bi-rads category 3 comparison: probably benign category after recall from screening before and after implementation of digital breast tomosynthesis. *Radiology*, 285(3):778–787, 2017.

Mouna Zouari Mehdi, Norhene Gargouri Ben Ayed, Alima Damak Masmoudi, Dorra Sellami, and Riadh Abid. An efficient microcalcifications detection based on dual spatial/spectral processing. *Multimedia Tools and Applications*, 76(11): 13047–13065, 2017.

Arianna Mencattini, Marcello Salmeri, Roberto Lojacono, Manuela Frigerio, and Federica Caselli. Mammographic images enhancement and denoising for breast cancer detection using dyadic wavelet processing. *IEEE transactions on instrumentation and measurement*, 57(7):1422–1430, 2008.

Gisela LG Menezes, Gonneke AO Winter-Warnars, Eva L Koekenbier, Emma J Groen, Helena M Verkooijen, and Ruud M Pijnappel. Simplifying breast imaging reporting and data system classification of mammograms with pure suspicious calcifications. *Journal of medical screening*, 25(2):82–87, 2018.

Jan-Jurre Mordang, Tim Janssen, Alessandro Bria, Thijs Kooi, Albert Gubern-Mérida, and Nico Karssemeijer. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In *International Workshop on Breast Imaging*, pages 35–42. Springer, 2016.

JJ Mordang, A Gubern-Mérida, A Bria, F Tortorella, RM Mann, MJM Broeders, GJ den Heeten, and N Karssemeijer. The importance of early detection of calcifications associated with breast cancer in screening. *Breast cancer research and treatment*, 167(2):451–458, 2018.

Chisako Muramatsu, Qiang Li, Robert A Schmidt, Junji Shiraishi, Kenji Suzuki, Gillian M Newstead, and Kunio Doi. Determination of subjective similarity for pairs of masses and pairs of clustered microcalcifications on mammograms: comparison of similarity ranking scores and absolute similarity ratings. *Medical physics*, 34(7): 2890–2895, 2007.

Mario Mustra and Mislav Grgic. Robust automatic breast and pectoral muscle segmentation from scanned mammograms. *Signal processing*, 93(10):2817–2827, 2013a.

Mario Mustra and Mislav Grgic. Robust automatic breast and pectoral muscle segmentation from scanned mammograms. *Signal processing*, 93(10):2817–2827, 2013b.

Jawad Nagi, Sameem Abdul Kareem, Farrukh Nagi, and Syed Khaleel Ahmed. Automated breast profile segmentation for roi detection using digital mammograms. In *2010 IEEE EMBS conference on biomedical engineering and sciences (IECBES)*, pages 87–92. IEEE, 2010a.

Jawad Nagi, Sameem Abdul Kareem, Farrukh Nagi, and Syed Khaleel Ahmed. Automated breast profile segmentation for roi detection using digital mammograms. In *2010 IEEE EMBS conference on biomedical engineering and sciences (IECBES)*, pages 87–92. IEEE, 2010b.

Ryohei Nakayama, Yoshikazu Uchiyama, Koji Yamamoto, Ryoji Watanabe, and Kiyoshi Namba. Computer-aided diagnosis scheme using a filter bank for detection of microcalcification clusters in mammograms. *IEEE Transactions on Biomedical engineering*, 53(2):273–283, 2006.

Constanza Pardo-Ramos and Ricardo Cendales-Duarte. *Incidencia, mortalidad y prevalencia de Cáncer en Colombia, 2007 - 2011*. Instituto Nacional de Cancerología - ESE, Colombia, primera edición edition, 2015.

Ah Young Park, Bo Kyoung Seo, Kyu Ran Cho, and Ok Hee Woo. The utility of micropure™ ultrasound technique in assessing grouped microcalcifications without a mass on mammography. *Journal of breast cancer*, 19(1):83–86, 2016.

Renbin Peng, Hao Chen, and Pramod K Varshney. Noise-enhanced detection of micro-calcifications in digital mammograms. *IEEE Journal of Selected Topics in Signal Processing*, 3(1):62–73, 2009.

Yonghong Peng, Bin Yao, and Jianmin Jiang. Knowledge-discovery incorporated evolutionary search for microcalcification detection in breast cancer diagnosis. *Artificial Intelligence in Medicine*, 37(1):43–53, 2006.

Hairong Qi and Wesley E Snyder. Content-based image retrieval in picture archiving and communications systems. *Journal of Digital Imaging*, 12(1):81–83, 1999.

David Raba, Arnau Oliver, Joan Martí, Marta Peracaula, and Joan Espunya. Breast segmentation with pectoral muscle suppression on digital mammograms. In *Iberian conference on pattern recognition and image analysis*, pages 471–478. Springer, 2005.

Poulami Raha, Radhika V Menon, and Indrajit Chakrabarti. Fully automated computer aided diagnosis system for classification of breast mass from ultrasound images. In

2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pages 48–51. IEEE, 2017.

M R Rakesh, B Ajeya, and A R Mohan. Hybrid Median Filter for Impulse Noise Removal of an Image in Image Restoration. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2 (10):5117–5124, 2013.

Andrik Rampun, Philip J Morrow, Bryan W Scotney, and John Winder. Fully automated breast boundary and pectoral muscle segmentation in mammograms. *Artificial intelligence in medicine*, 79:28–41, 2017.

Jinchang Ren. Ann vs. svm: Which one performs better in classification of mccs in mammogram imaging. *Knowledge-Based Systems*, 26:144–153, 2012.

A Retico, Pasquale Delogu, MARIA EVELINA Fantacci, A Preite Martinez, A Stefanini, and A Tata. A scalable computer-aided detection system for microcalcification cluster identification in a pan-european distributed database of mammograms. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 569(2):601–605, 2006.

Christian Robert. Machine learning, a probabilistic perspective, 2014.

A Rosebrock. The complete guide to building an image search engine with python and opencv (2014), 2017.

Rahimeh Rouhi, Mehdi Jafari, Shohreh Kasaei, and Peiman Keshavarzian. Benign and malignant breast tumors classification based on region growing and cnn segmentation. *Expert Systems with Applications*, 42(3):990–1002, 2015.

Yong Rui. Human perception subjectivity and relevance feedback in multimedia information retrieval. *Proc. of IS&T and SPIE Storage and Retrieval of Image and Video Databases, San Jose, CA, Jan, 1998*, 1998.

Yong Rui and Thomas S Huang. A novel relevance feedback technique in image retrieval. In *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*, pages 67–70, 1999.

Paolo Russo. *Handbook of X-ray imaging: physics and technology*. CRC Press, 2017.

Alvaro Andres Sandino Garzón et al. Detección de microcalcificaciones mamarias agrupadas. Master's thesis, Universidad Distrital Francisco José de Caldas, 2018.

Ni Larasati Kartika Sari, Prawito Prajitno, Lukmanda Evan Lubis, and Djarwani Soeharso Soejoko. Computer aided diagnosis (cad) for mammography with markov random field method with simulated annealing optimization. *Journal of Medical Physics and Biophysics*, 4(1):85–94, 2017.

Vikas Kumar Saubhagya, Asha Rani, and Vijander Singh. Ann based detection of breast cancer in mammograph images. In *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, pages 1–6. IEEE, 2016.

Kai Scherer, Eva Braig, Sebastian Ehn, Jonathan Schock, Johannes Wolf, Lorenz Birnbacher, Michael Chabior, Julia Herzen, Doris Mayr, Susanne Grandl, et al. Improved diagnostics by assessing the micromorphology of breast calcifications via x-ray dark-field radiography. *Scientific reports*, 6:36991, 2016.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. Introduction to information retrieval. In *Proceedings of the international communication of association for computing machinery conference*, volume 4, 2008.

Juan Shan, S Kaisar Alam, Brian Garra, Yingtao Zhang, and Tahira Ahmed. Computer-aided diagnosis for breast ultrasound using computerized bi-rads features and machine learning methods. *Ultrasound in medicine & biology*, 42(4):980–988, 2016.

Ayush Shrivastava, Arpit Chaudhary, Devang Kulshreshtha, Vibhav Prakash Singh, and Rajeev Srivastava. Automated digital mammogram segmentation using dispersed region growing and sliding window algorithm. In *2017 2nd international conference on image, vision and computing (ICIVC)*, pages 366–370. IEEE, 2017.

Neeraj Shrivastava and Jyoti Bharti. Breast tumor detection in digital mammogram based on efficient seed region growing segmentation. *IETE Journal of Research*, pages 1–13, 2020.

Chi-Ren Shyu, Carla E Brodley, Avinash C Kak, Akio Kosaka, Alex M Aisen, and Lynn S Broderick. Assert: A physician-in-the-loop content-based retrieval system for hrct image databases. *Computer Vision and Image Understanding*, 75(1-2): 111–132, 1999.

Birmohan Singh and Manpreet Kaur. An approach for classification of malignant and benign microcalcification clusters. *Sādhanā*, 43(3):39, 2018.

Aakriti Singla and Anand Sharma. Physical access system security of iot devices using machine learning techniques. *Available at SSRN 3356785*, 2019.

Magny S.J., Shikhman R., and Keppke A.L. Breast, imaging, reporting and data system (bi rads). *StatPearls [Internet], StatPearls Publishing*, NA, 2020.

Marijeta Slavković-Ilić, Ana Gavrovska, Milan Milivojević, Irini Reljin, and Branimir Reljin. Breast region segmentation and pectoral muscle removal in mammograms. *Telfor Journal*, 8(1):50–55, 2016.

Chris Solomon and Toby Breckon. *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons, 2011.

Khaoula Belhaj Soulami, Mohamed Nabil Saidi, Bouchra Honnit, Chaimae Anibou, and Ahmed Tamtaoui. Detection of breast abnormalities in digital mammograms using the electromagnetism-like algorithm. *Multimedia Tools and Applications*, 78 (10):12835–12863, 2019.

Yajie Sun, R Janer, Jasjit Suri, Zhen Ye, and RM Rangayyan. Effect of adaptive-neighborhood contrast enhancement on the extraction of the breast skin-line in mammograms. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 3475–3478. IEEE, 2006a.

Yajie Sun, Jasjit S Suri, JE Leo Desautels, and Rangaraj M Rangayyan. A new approach for breast skin-line estimation in mammograms. *Pattern analysis and applications*, 9(1):34, 2006b.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Rong Tan, Ying Xiao, Qi Tang, Ying Zhang, Hui Chen, and Xiancheng Fan. The diagnostic value of micropure imaging in breast suspicious microcalcificaion. *Academic radiology*, 22(11):1338–1343, 2015.

Fabienne Thibault, Martine Meunier, Jerzy Klijanienko, Carl El Khoury, Claude Nos, Anne Vincent-Salomon, Bernard Asselain, and Sylvia Neuenschwander. Diagnostic accuracy of sonography and combined sonographic assessment and sonographically guided cytology in nonpalpable solid breast lesions. *Journal of clinical ultrasound*, 28(8):387–398, 2000.

Alain Tiedeu, Christian Daul, Aude Kentsop, Pierre Graebling, and Didier Wolf. Texture-based analysis of clustered microcalcifications detected on mammograms. *Digital Signal Processing*, 22(1):124–132, 2012.

Lazaros Tsochatzidis, Konstantinos Zagoris, Nikolaos Arikidis, Anna Karahaliou, Lena Costaridou, and Ioannis Pratikakis. Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach. *Pattern Recognition*, 71:106–117, 2017.

Paula van Luijt, Jacques Fracheboud, Eveline Heijnsdijk, Gerard den Heeten, and Harry de Koning. Performance during the transition to digital mammography. observations in 6 million screens. *Overdiagnosis in the Dutch and Norwegian breast cancer screening program*, 49(16):29, 2013.

PS Vikhe and VR Thool. Contrast enhancement in mammograms using homomorphic filter technique. In *2016 International Conference on Signal and Information Processing (IConSIP)*, pages 1–5. IEEE, 2016.

Juan Wang and Yongyi Yang. A context-sensitive deep learning approach for microcalcification detection in mammograms. *Pattern recognition*, 78:12–22, 2018.

Xingwei Wang, Lihua Li, Wei Liu, Weidong Xu, Dror Lederman, and Bin Zheng. An interactive system for computer-aided diagnosis of breast masses. *Journal of digital imaging*, 25(5):570–579, 2012.

Xuechuan Wang and Kuldip K Paliwal. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern recognition*, 36(10):2429–2439, 2003.

Chia-Hung Wei, Sherry Y Chen, and Xiaohui Liu. Mammogram retrieval on similar mass lesions. *Computer methods and programs in biomedicine*, 106(3):234–248, 2012.

Liyang Wei, Yongyi Yang, Robert M Nishikawa, Miles N Wernick, and Alexandra Edwards. Relevance vector machine for automatic detection of clustered microcalcifications. *IEEE transactions on medical imaging*, 24(10):1278–1285, 2005.

Stefanie Weigel, Hans W Hense, Jan Heidrich, Shoma Berkemeyer, Walter Heindel, and Oliver Heidinger. Digital mammography screening: does age influence the detection rates of low-, intermediate-, and high-grade ductal carcinoma in situ? *Radiology*, 278(3):707–713, 2015.

Weiying Xie, Yunsong Li, and Yide Ma. Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing*, 173:930–941, 2016.

Erhu Zhang, Fan Wang, Yongchao Li, and Xiaonan Bai. Automatic detection of microcalcifications using mathematical morphology and a support vector machine. *Bio-medical materials and engineering*, 24(1):53–59, 2014.

Dongming Zhao. Rule-based morphological feature extraction of microcalcifications in mammograms. In *Biomedical Image Processing and Biomedical Visualization*, volume 1905, pages 702–715. International Society for Optics and Photonics, 1993.

David Augusto Cárdenas Peña, (Ibagué, Tolima, Colombia,1987)

Doctor en Ingeniería e Ingeniero Electrónico de la Universidad Nacional de Colombia.

Docente auxiliar transitorio de tiempo completo, facultad de ingenierías.

Ha publicado artículos en revistas especializadas nacionales e internacionales.

Vinculado al grupo de investigación en Automática.

dcardenasp@utp.edu.co

Santiago Marín Mejía
(Pereira, Risaralda, Colombia, 1992).

Magister en Innovación Empresarial e Informática de la Università degli Studi di Salerno y en Ingeniería de Sistemas y Computación de la Universidad Tecnológica de Pereira. Ingeniero Industrial de la Universidad Tecnológica de Pereira.

Ha publicado artículos en revistas especializadas nacionales e internacionales.

santiagomarin92@utp.edu.co

La Editorial de la Universidad Tecnológica de Pereira tiene como política la divulgación del saber científico, técnico y humanístico para fomentar la cultura escrita a través de libros y revistas científicas especializadas.

Las colecciones de este proyecto son: Trabajos de Investigación, Ensayos, Textos Académicos y Tesis Laureadas.

Este libro pertenece a la Colección Trabajos de Investigación.

Este libro presenta los alcances y resultados del proyecto de investigación "Prototipo de un sistema de recuperación de información por contenido orientado a la clasificación de grupos de microcalcificaciones en mamografías" desarrollado por el Grupo de Investigación en Automática y por el Grupo de Investigación en Análisis de Datos y Sociología Computacional de la Facultad de Ingenierías de la Universidad Tecnológica de Pereira. En este libro se aborda el desarrollo de una metodología de recuperación de información cuyo fin último es asistir a los especialistas médicos en el análisis de imágenes de mamografías digitales y en el posterior descubrimiento de patrones que puedan ser indicadores de la existencia de microcalcificaciones en los tejidos mamarios.

El desarrollo que se presenta en este libro, el cual está enmarcado dentro de las ciencias de la computación y, específicamente, en el área del aprendizaje de máquina, podría coadyuvar a los especialistas en la detección temprana del cáncer de mama, permitiéndoles acceder a través de un sistema inteligente de recuperación de información a datos históricos de diagnósticos confirmados que estén estrechamente relacionados con las características puntuales del caso bajo estudio. En este libro se presentan en detalle cada una de las técnicas utilizadas en los módulos que componen la metodología, además de su funcionamiento y de los resultados obtenidos sobre bases de datos previamente etiquetadas.

Facultad de Ingenierías
Colección Trabajos de Investigación

eISBN 978-958-722-517-4